Florida Assessments for Instruction in Reading, Aligned to the Language Arts Florida Standards

FAIR – FS

Grades 3 through 12

Technical Manual

Barbara R. Foorman, Ph.D.

Yaacov Petscher, Ph.D.

Chris Schatschneider, Ph.D.

Florida Center for Reading Research
Florida State University

Version 1.2

3/26/2015

Acknowledgements

The items, dynamic flow, computer-adaptive task flow and placement algorithms, creation of the development application, and psychometric work for this component skills battery (called the Florida Center for Reading Research Reading Assessment; FRA) were funded by grants from the Institute of Education Sciences (IES) to Florida State University:

Institute of Education Sciences, USDOE ($4,447,990), entitled "Assessing Reading for Understanding: A Theory-Based, Developmental Approach," subcontract to the Educational Testing Service for five years (R305F100005), 7/1/10-6/30/15 (Foorman, PI on subcontract; Petscher and Schatschneider, Co-Is).

Institute of Education Sciences, USDOE (R305A100301; $1,499,741), entitled "Measuring Reading Progress in Struggling Adolescents," awarded for four years, 3/1/10-2/28/14. (Foorman, PI; Petscher and Schatschneider, Co-Is).

The Florida State University licensed the FRA to the Florida Department of Education at no cost in perpetuity in 2012. We would like to acknowledge the following individuals for their leadership in executing the work funded by the above two IES grants: Dr. Adrea Truckenmiller, Karl Hook, and Nathan Day. We also would like to thank the numerous school districts administrators and teachers who participated in the research funded by these two grants.

For the purpose of this technical manual, we refer to the portions of the FRA licensed to the Florida Department of Education as the Florida Assessments for Instruction in Reading, Aligned to the Language Art Florida Standards (FAIR-FS). Components of the FAIR-FS in grades 3-12 which are owned by the Florida Department of Education (e.g., ORT) are described in a separate technical manual.

Table of Contents

Introduction

The first question to ask when designing an assessment of reading and language skills is what predicts success in comprehending written language, that is, success in word reading and in reading comprehension? We are fortunate to have several consensus documents that review decades of literature about what predicts reading success (NRC, 1998; NICHD, 2000; NIFL, 2008; Rand, 2002; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001).

Mastering the Alphabetic Principle

What matters the most to success in reading words in an alphabetic orthography such as English is mastering the alphabetic principle, the insight that speech can be segmented into discrete units (i.e., phonemes) that map onto orthographic (i.e., graphemic) units (Ehri, Nunes, Willows, et al., 2001; Rayner et al., 2001). Oral language is acquired largely in a natural manner within a hearing/speaking community; however, written language is not acquired naturally because the graphemes and their relation to phonological units in speech are invented and must be taught by literate members of the community. The various writing systems (i.e., orthographies) of the world vary in the transparency of the sound-symbol relation. Among alphabetic orthographies, the Finnish orthography is highly transparent: phonemes in speech relate to graphemes in print (i.e., spelling) in a highly consistent one-to-one manner and graphemes in print relate to phonemes in speech (i.e., decoding) in a highly consistent one-to-one manner. Thus, learning to spell and read Finnish is relatively easy. English, however, is a more opaque orthography. Phonemes often relate to graphemes in an inconsistent manner and graphemes relate to phonemes in yet a different inconsistent manner. For example, if we hear the "long sound of *a*" we can think of words with many different vowel spellings, such as *crate*, *brain*, *hay*, *they*, *maybe*, *eight*, *great*, *vein*. If we see the orthographic unit *–ough*, we may struggle with the various pronunciations of *cough*, *tough*, *though*, *bough*. The good news is that 69% of monosyllabic English words—those Anglo-Saxon words most used in beginning reading instruction—are consistent in their letter to pronunciation mapping (Ziegler, Stone, & Jacobs, 1997). Most of the rest can be learned with grapheme-phoneme correspondence rules (i.e., phonics), with only a small percentage of words being so irregular in their letter-sound relations that they should be taught as sight words (Ehri, Nunes, Stahl, & Willows, 2001; Foorman & Connor, 2011).

In the FAIR-FS, the alphabetic principle is assessed in grades K-2 with individually-administered tasks that measure letter-sound knowledge, phonological awareness, ability to link sounds to letters, word reading, word building, and spelling tasks. All Screening tasks are computer-adaptive, with 5 items presented at grade level before the system adapts to easier or more difficult items based on student ability, and with the teacher scoring the responses as correct or incorrect. In kindergarten, the Screening tasks consist of asking students: 1) to name the sound of letters presented on the computer monitor; 2) to blend sounds pronounced by the computer into words; and, 3) at the end of the year, to read simple words presented on the computer monitor. In grades 1 and 2 the Screening task consists of a computer-adaptive word list where students pronounce a word presented on the computer monitor. In grade 2,

students use the keyboard to spell the word pronounced by the computer and used in a sentence. Score reports include students' misspellings and a guide for analyzing errors is provided in the administration manual. If K-2 students' performance on the Screening tasks is predicted to be below the 40^{th} percentile on the Stanford Achievement Tests (SESAT Word Reading in kindergarten and reading comprehension in grades 1-2), they go on to take the Diagnostic tasks, which are computer-administered but scored on a mastery criteria. The skills progress from print awareness, to 26 letter names and 29 letter-sounds (including three digraphs), to deleting initial and final sounds and matching them to the correct letters, to phonological blending and deletion, to building words in CVC, CVCe, CVCC, and CCVC patterns, to reading multisyllabic words.

In grades 3-12, alphabetic skills are measured with a word recognition task. In this computer-adaptive task, three words are presented on the computer monitor and students must select the word that best matches the word pronounced by the computer. About 10% of target words are nonsense words so that phonological decoding skills are tapped. When the target is a real word, distractors tap orthographic knowledge. For example, a distractor for "prerogative" might be *perogative*. By tapping orthographic knowledge in this task, the quality of a student's lexical representation for a printed word is assessed. The more complete and accurate the lexical representation of a word is, the more efficient the student's word recognition and reading comprehension (Perfetti & Stafura, 2014).

Comprehending Written Language (better known as Reading Comprehension)

Knowledge of word meanings*.* Mastering the alphabetic principle is a necessary but not sufficient condition for understanding written text. We may be able to pronounce printed words, but if we don't know their meaning our comprehension of the text is likely to be impeded. Hence, our knowledge of word meanings is crucial to comprehending what we read. Grasping the meaning of a word is more than knowing its definition in a particular passage. Knowing the meaning of a word means knowing its full lexical entry in a dictionary: pronunciation, spelling, multiple meanings in a variety of contexts, synonyms, antonyms, idiomatic use, related words, etymology, and morphological structure. For example, a dictionary entry for the word *exacerbate* says that it is a verb meaning: 1) to increase the severity, bitterness, or violence of (disease, ill feeling, etc.); aggravate or 2) to embitter the feelings of (a person); irritate; exasperate (e.g., foolish words that only exacerbated the quarrel). It comes from the Latin word *exacerbātus* (the past participle of *exacerbāre*: to *exasperate*, *provoke*), equivalent to *ex* + *acerbatus* (*acerbate*). Synonyms are: *intensify*, *inflame*, *worsen*, *embitter*. Antonyms are: *relieve*, *sooth*, *alleviate*, *assuage*. Idiomatic equivalents are: add fuel to the flame, fan the flames, feed the fire, or pour oil on the fire. The more a reader knows about the meaning of a word like *exacerbate*, the greater the lexical quality the reader has and the more likely the reader will be able to recognize the word quickly in text, with full comprehension of its meaning (Perfetti & Stafura, 2014).

In the grades 3-12 FAIR-FS, knowledge of word meanings is measured by a Vocabulary Knowledge Task that taps morphological awareness. In the Vocabulary Knowledge Task, the student reads a sentence that has a missing word. The student selects among three words the one that best completes the sentence. The distractors and target vary in their morphological structure (i.e., prefixes or suffixes

consisting of inflectional morphemes or derivational morphemes). It is relatively easy to read derived words that are pronounced similarly to their base (e.g., *reason*, *reasonable*). Words that contain a phonological shift (e.g., *vine*, *vineyard*) or an orthographic shift (e.g., *pity*, *piteous*) are harder to read, and words that contain both a phonological and an orthographic shift (e.g., *theory*, *theoretical*) are the hardest of all (Carlisle & Stone, 2005). The Vocabulary Knowledge Task in the FAIR-FS explained 2%-9% unique variance beyond prior reading comprehension, text reading efficiency, and spelling in predicting spring reading comprehension (Foorman, Petscher, & Bishop, 2012) and, by doing so, addresses aspects of language critical to understanding written language, language often called *academic language* because it is found in books and at school but not in informal conversations at home or outside school. Part of academic language is *inferential language* or *decontextualized language*, which allows speakers or writers to go beyond the present context and to predict, hypothesize, compare and contrast, and reason about events (e.g., an upcoming *referendum*) or abstract concepts (e.g., *photosynthesis*, *gravity*). Examples of words that signal such inferential or decontextualized language are *describe*, *analyze*, *hypothesize*.

Syntactic awareness*.* In addition to understanding word meanings, another important aspect of academic language is syntactic awareness. Syntax or grammar refers to the rules that govern how words are ordered to make meaningful sentences. Children typically acquire these rules in their native language prior to formal schooling. However, learning to apply these rules to reading and writing is a goal of formal schooling and takes years of instruction and practice. In the grades 3-12 FAIR-FS, there is a diagnostic task called Syntactic Knowledge Task (SKT). In this task the student listens to a sentence that is missing a word and selects the best word from a dropdown menu to complete the sentence. The words are verbs, pronouns, or connectives. Connectives are words that represent causal (e.g., *because*), temporal (e.g., *when*), logical (e.g., *if-then*), additive (e.g., *in addition*), or adversative (e.g., *although*) relations and are important linguistic devices for linking ideas and information within and across sentences. They link back to information already read through pronoun reference (anaphora) or repetition of nouns and verbs and provide clues to future meaning (e.g., *therefore*, *nonetheless*). Knowledge of the meaning and use of connectives is an important aid to comprehension (Cain & Nash, 2011; Crosson & Lesaux, 2013).

Reading comprehension*.* If a student can read and understand the meanings of printed words and sentences, then comprehending text should not be difficult, given the emphasis above on achieving the alphabetic principle, lexical quality, and syntactic awareness. Individual differences in readers' background knowledge, motivation, and memory and attention will create variability in word recognition skills, vocabulary knowledge, and syntactic awareness and this variability, in turn, will create variability in reading comprehension. Furthermore, genre differences—informational or literary text—may interact with reader skills to affect reading comprehension. For example, some students may have better inferential language skills so critical to comprehending informational text; other students may have better narrative language skills of discerning story structure and character motivation and, therefore, be good comprehenders of literary text. Because reading comprehension is affected by the interactions of variables related to reader and text characteristics (RAND, 2002), tests of reading

comprehension typically consist of informational *and* literary passages and provide as much relevant background information within the passage as possible.

States' reading comprehension tests typically have questions written to their state standards. One challenge for these tests are the trade-offs between coverage of the standards, time, and reliability. Typically, one should strive for about 15 items per standard. If a state has 14 standards per grade, then 210 questions would be needed to reliably cover the standards. If 7-9 questions are written for each passage, then students would need to read 23-30 passages, which would take them about 10 days. Most states prioritize testing the superordinate standards in order to reduce the testing time to 7 passages or so over two days. A limitation of many standards-based tests is their sole focus on grade-level proficiency. Students are given only grade-level passages; therefore, students who read below grade level tend to guess and students who read above grade level are not challenged. In both cases, no information about their actual reading ability is obtained. Furthermore, when the grade level of passages is determined by readability formulae or by qualitative ratings, the precision is not at a particular grade but rather within grade bands of two to three grades (e.g., upper elementary, middle school, high school; Foorman, 2009; Nelson, Perfetti, Liben, & Liben, 2012).

The FAIR-FS Reading Comprehension task in grades 3-12 avoids the problems with precision and efficiency noted above by being a computer-adaptive test. Students are placed into their first reading comprehension passage based on their ability on the computer-adaptive Word Recognition and Vocabulary Knowledge Tasks—which take 2-3 minutes each. The student reads the passage and answers the 7-9 multiple choice questions. Subsequent passage placement is based on relations among student ability, standard error, and discrimination parameters from a 2-parameter logistic item response theory (IRT) model. Students continue to receive passages until a precise estimate of reading comprehension is achieved (i.e., reliability >.80). In the FAIR-FS, students receive 1-3 passages in about 10-30 minutes. Given that the two Screening tasks and one Diagnostic task take, on average, 11 minutes, the entire 3-12 battery easily fits into a 45-minute class period. During the 2013-2014 implementation study in Pinellas County, reliability on the Reading Comprehension task was above .80 for 93 percent of students and above .90 for 54 percent of students.

Individual tasks in the FAIR-FS yield two score types—percentile ranks and ability scores. The ability score is used to measure growth and can be displayed against grade-level percentile ranks to communicate the important point that students are improving across the year even though they are performing far below or above grade-level peers.

Summary of FAIR-FS Constructs and Tasks

The FAIR-FS consists of computer-adaptive reading comprehension and oral language screening tasks that provide measures to track growth over time, as well as a Probability of Literacy Success (PLS) linked to grade-level performance (i.e., the 40th percentile) on the reading comprehension subtest of the Stanford Achievement Test (SAT-10) in the 2014-2015 school year and will predict to the Florida Standards Assessment once those data are available. Thus, the FAIR-FS provides universal screening and

diagnostic tasks in a precise and efficient computer-adaptive framework with psychometrics and norms derived from large samples of Florida K-12 students representative of Florida demographics. By including Vocabulary Knowledge and Syntax Knowledge Tasks, the FAIR-FS has excellent construct coverage of oral language, which has been shown to account for the vast majority (i.e., 72%-96%, with a median of 87%) of individual differences in reading comprehension in grades 4-10 (Foorman, Koon, Petscher, Mitchell, & Truckenmiller, 2015) and comparable variance to decoding fluency in grades 1-2 (Foorman, Herrera, Petscher, Mitchell, & Truckenmiller, 2015).

Description of the Tasks in the FAIR-FS

Item development. Item development was broadly based on the empirical theories regarding reading development described above. Retention for specific items was based principally on the statistical properties of the items and is detailed in the Description of Method section. Items were originally written and reviewed by a team of experienced educators with advanced degrees in education, communication, and psychology. Item writers generally wrote to late elementary, middle, and high school students using vocabulary and text complexity that the writers had experienced in typical curricula and materials targeted to those age groups. Item writers created a variety of items that they considered to be easy, moderate, and difficult for the range of students. Writers were asked to provide a larger number of easier and moderate items. Given that screening assessments are more commonly given to lower performing students and those students are assessed more frequently, the item bank needed to have a large number of easy and moderate items so that there were enough items in the item bank that students did not have to see the same items each year. Each item was reviewed by at least three other members of the review team for errors and appropriateness. All items in the Reading Comprehension task were aligned with a standard from the Common Core State Standards. As part of the FAIR-FS contract with the Florida Department of Education, members of the Just Read, Florida! office also reviewed the Reading Comprehension Task passages and questions specifically for alignment to the Language Arts Florida Standards.

Target words for the WRT and VKT tasks were based on pilot work with a small group of students and printed word frequency (Zeno, Ivens, Millard, & Duvvuri, 1995). A rough estimate of the range in difficulty of the sentences in the VKT and SKT tasks was obtained through use of the Flesch-Kincaid grade-level readability formula.

Passages and items in the Reading Comprehension Task were written to address the Language Arts Florida Standards in three strands (Reading Informational Text, Reading Literary Text, and Language). Items writers also reviewed publicly available examples from the Florida Comprehensive Assessment Test, the Partnership for Assessment of Readiness for College and Careers and the SmarterBalanced Consortium. The range of text complexity of the passages was evaluated for a variety of freely available quantitative measures (i.e., Lexile, Flesch-Kincaid, Pearson Maturity Metric, Text Evaluator, ATOS, and Degrees of Reading Power) and the qualitative rating guide from Appendix A of the Common Core State Standards. The passages in elementary grades were originally written to be evenly split between literary and informational passages. The passage and item difficulty was ultimately determined by the

normative sample's performance on the task, so the resulting item bank is split 42% literary passages and 58% informational. Since the goal of this assessment is to cover the range of student ability as opposed to equally addressing all standards, the guidelines for item creation on the Reading Comprehension task was to make 30% of the items focused on vocabulary and 70% of the items focused on explicit and inferential comprehension questions. The comprehension items for elementary aged students were split evenly between explicit and implicit questions with the percentage favoring implicit questions at the upper grade levels.

Word Recognition Task (WRT). In the Word Recognition Task, the student listens to a word pronounced by the computer. The computer monitor displays a drop-down menu with the correctly spelled word and two distractors that are spelled incorrectly. The student may replay the audio for the word up to three times. The student has unlimited time to respond to each item. The item bank contains 274 available items and includes real words and some non-words. The range of possible theta scores in the WRT is -3.88 to 3.85. This range corresponds to an ability score range of 112 to 885.

Vocabulary Knowledge Task (VKT). Each item in the Vocabulary Knowledge Task consists of one sentence with a word missing. The missing word is replaced with a choice of three morphologically related words. The student selects the word that best completes the sentence. There are 374 items available. The student has unlimited time to respond to each item. The range of possible theta scores in the VKT is -2.55 to 3.59. This range corresponds to an ability score range of 245 to 859.

Reading Comprehension (RC). The Reading Comprehension task consists of passages that are between 200 and 1300 words in length. Each passage has between 7 and 9 multiple choice questions. Each question has one correct response and three distractors. All questions associated with the passage are displayed at the same time and the passage is also available on the computer monitor. Each question has an individual item difficulty and discrimination value. Each set of 7 to 9 questions has an average item difficulty, which is used to determine which set of questions (and associated passage) is administered to the student next. The Reading Comprehension task ends when a reliable score has been reached (i.e., the standard error is less than 0.316) or the student has responded to three sets of questions. The initial set of questions administered to a student is determined by a formula that includes the student's score on the WRT and the VKT. The computer will automatically log out students after 15 minutes of inactivity; otherwise, students have an unlimited amount of time to read the passage and respond to questions. There are a total of 139 sets of questions associated with passages available in the grades 3-12 FAIR-FS. The range of possible theta scores in the RCT is -2.80 to 5.24. This range corresponds to an ability score range of 220 to 1024.

Syntactic Knowledge Task (SKT). In the Syntactic Knowledge Task, the student listens to a sentence or sentences read by the computer that is missing one word. The computer monitor also displays the sentence(s) for the student to read along. The missing word(s) in the sentence(s) is replaced by a dropdown box with the correct word or phrase and two distractors. There are a total of 240 items available. Some items require a student to select the correct connective word, the correct pronoun

reference, or the correct verb that creates appropriate subject-verb agreement. The range of possible theta scores in the SKT is -3.08 to 3.34. This range corresponds to an ability score range of 192 to 834.

Task Administration. In grades 3 through 12, the FAIR-FS consists of four computer-adaptive tasks that each provide unique information regarding a student's literacy skills. Each of the tasks below, except for Reading Comprehension, have four stop rules that determine when administration of each task is complete[1].

1. A reliable estimate of the student's abilities is reached (i.e., standard error is less than 0.316).
2. The student has responded to 30 items.
3. The student responds correctly to all of the first 8 items.
4. The student responds incorrectly to all of the first 8 items.

At subsequent administrations of the tasks within the same school year, the student's prior score on that task determines the initial set of items administered to the student at that administration period.

The tasks in the FAIR-FS can be used as a highly efficient diagnostic tool due to the utilization of computer adaptive functionality. Computer administration allows for large groups of students to be assessed at once with a high degree of standardization. Adaptability in the items allows for a highly reliable score to be reached sooner and decreases the amount of time needed for each task. Although educators are most concerned with students' abilities in reading comprehension, it is a complex skill that takes significant amounts of time to assess (due to close reading of extended text) and poor performance does not necessarily signal which component skills of reading to target for instruction. The FAIR-FS efficiently assesses multiple research-based component skills of reading comprehension to help teachers diagnose skill weaknesses and target instruction. During the implementation study, more than 98% of students reached a highly reliable score (marginal reliability above .80) by taking an average of only 20 items on the WRT, 9 items on the VKT, and 18 items on the SKT. Table 1 provides a description of the efficiency of each task. The increase in efficiency allows for more tasks to be administered to achieve a more complete diagnostic profile for a student. For example, in the implementation study 84% of students in grades 3 through 12 completed all four of the computer-adaptive tasks within one class period (i.e., 45 minutes).

1 The stop rules for reading comprehension are a maximum of three passages or a reliable estimate of the student's ability (i.e., standard error < .316).

Table 1

Task Efficiency

	Word Recognition Task	Vocabulary Knowledge Task	Syntactic Knowledge Task	Reading Comprehension Task	
	Number of items			Passages administered	% students
mean	20	9	17	1 passage	9.7%
median	19	8	16	2 passages	22.7%
administered 30 items	31%	2%	15%	3 passages	67.6%
	Reliability				
marginal reliability coefficient	0.93	0.91	0.93		0.94
Cronbach's alpha ≥ .9	82%	98%	87%		54%
Cronbach's alpha ≥ .8	98%	99%	99%		93%
	Time (minutes : seconds)				
mean	3:04	2:06	3:54		NA*
median	2:36	1:40	3:30		NA*
directions time	0:42	0:24	0:35		0:15

*The mean and median values for amount of time spent on the Reading Comprehension Task are not available due to the nature of the task.

Description of Method

Item tryout and validation work with the above tasks occurred from 2010-2015 through the funding provided by two IES grants (see Acknowledgements). Once item writers had written items for each task, tasks were piloted with students in grades 3-12. Results from Item Response Theory (IRT) analyses were evaluated and in several cases items were deleted or more difficult items were written and further field trials were conducted. A large-scale linking study was conducted during the Spring of 2013 with approximately 45,000 students in grades 3 through grade 12 in two districts in Florida. Outcome data consisted of well-known standardized measures of reading comprehension (Gates-MacGinitie and the SAT-10). Item response and differential item function analyses were conducted. Parameters derived from these analyses are used in the look-up tables in the computer-adaptive system.

Item Response Theory

Data for the grades 3-12 FAIR-FS were analyzed using Item Response Theory (IRT). Traditional testing and analysis of items involves estimating the difficulty of the item (based on the percentage of respondents correctly answering the item) as well as discrimination (how well individual items relate to overall test performance). This falls into the realm of measurement known as classical test theory (CTT). While such practices are commonplace in assessment development, IRT holds several advantages over CTT. When using CTT, the difficulty of an item depends on the group of individuals on which the data were collected. This means that if a sample has more students that perform at an above-average level, the easier the items will appear; but if the sample has more below-average performers, the items will appear to be more difficult. Similarly, the more that students differ in their ability, the more likely the discrimination of the items will be high; the more that the students are similar in their ability, the lower the discrimination will be. One could correctly infer that scores from a CTT approach are entirely dependent on the makeup of the sample on which the items are tested.

The benefits of IRT are such that: 1) the difficulty, discrimination, and pseudo-guessing parameters are not dependent on the group(s) from which they were initially estimated; 2) scores describing students' ability are not related to the difficulty of the test; 3) shorter tests can be created that are more reliable than a longer test; and, 4) item statistics and the ability of students are reported on the same scale.

Item Difficulty. The difficulty of an item has traditionally been described for many tests as a "p-value", which corresponds to the percent of respondents correctly answering an item. Values from this perspective range from 0% to 100% with high values indicating easier items and low values indicating hard items. Item difficulty in an IRT model does not represent proportion correct, but is rather represented as estimates along a continuum of -3.0 to +3.0. Figure 1 demonstrates a sample item characteristic curve which describes item properties from IRT. Along the x-axis is the ability of the individual, denoted by theta. As previously mentioned, the ability of students and item statistics are reported on the same scale. Thus, the x-axis is a simultaneous representation of student ability and item difficulty. Negative values along the x-axis will indicate that items are easier, while positive values describe harder items. Pertaining to students, negative values describe individuals who perform below average, while positive values identify students who perform above average. A value of zero for both students and items reflects average level of either ability or difficulty.

Along the y-axis is the probability of a correct response, which varies across the level of difficulty. Item difficulty is defined as the value on the x-axis at which the probability of correctly endorsing the item is 0.50. As demonstrated for the sample item in Figure 1, the difficulty of this item would be 0.0. Item characteristic curves are graphical representations generated for each item that allow the user to see how the probability of getting the item correct changes for different levels of the x-axis. Students with an ability of -3.0 would have an approximate 0.01 chance of getting the item correct, while students with an ability of 3.0 would have a nearly 99% chance of getting an item correct.

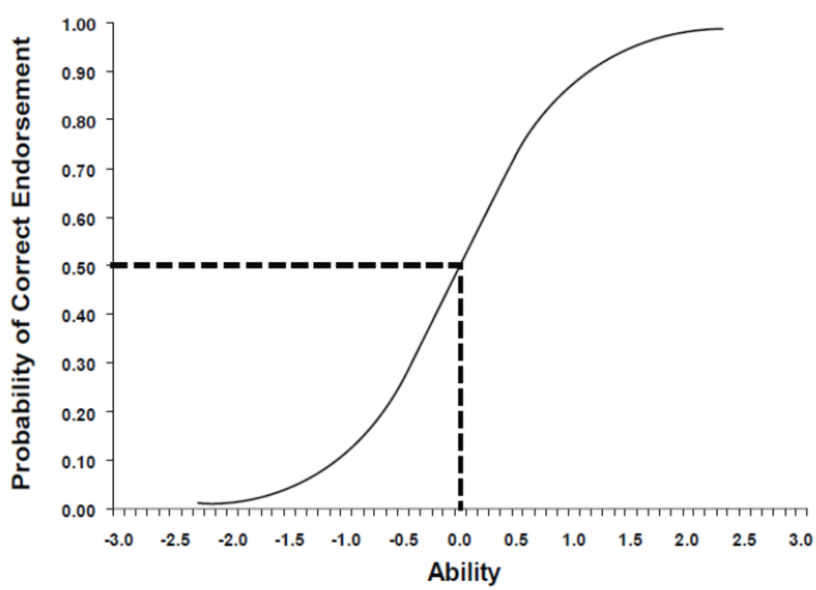


Figure 1. Sample Item Characteristic Curve

Item Discrimination. Item Discrimination is related to the relationship between how a student responds to an item and their subsequent performance on the rest of a test. In IRT it describes the extent to which an item can differentiate the probability of correctly endorsing an item across the range of ability (i.e., -3.0 to +3.0). Figure 2 provides an example of how discrimination operates in the IRT framework. For all three items presented in Figure 2, the difficulty has been held constant at 0.0, while the discriminations are variable. The dashed line (Item 1) shows an item with strong discrimination, the solid line (Item 2) represents an item with acceptable discrimination, and the dotted line (Item 3) is indicative of an item that does not discriminate. It is observed that for Item 3, regardless of the level of ability for a student, the probability of getting the item right is the same. Both high ability students and low ability students have the same chance of doing well on this item. Item 1 demonstrates that as the x-axis increases, the probability of getting the item correct changes as well. Notice that small changes between -1.0 and +1.0 on the x-axis result in large changes on the y-axis. This indicates that the item discriminates well among students, and that individuals with higher ability have a greater probability of getting the item correct. Item 2 shows that while an increase in ability produces an increase in the probability of a correct response, the increase is not as large as is observed for Item 1, and is thus a poorer discriminating item.

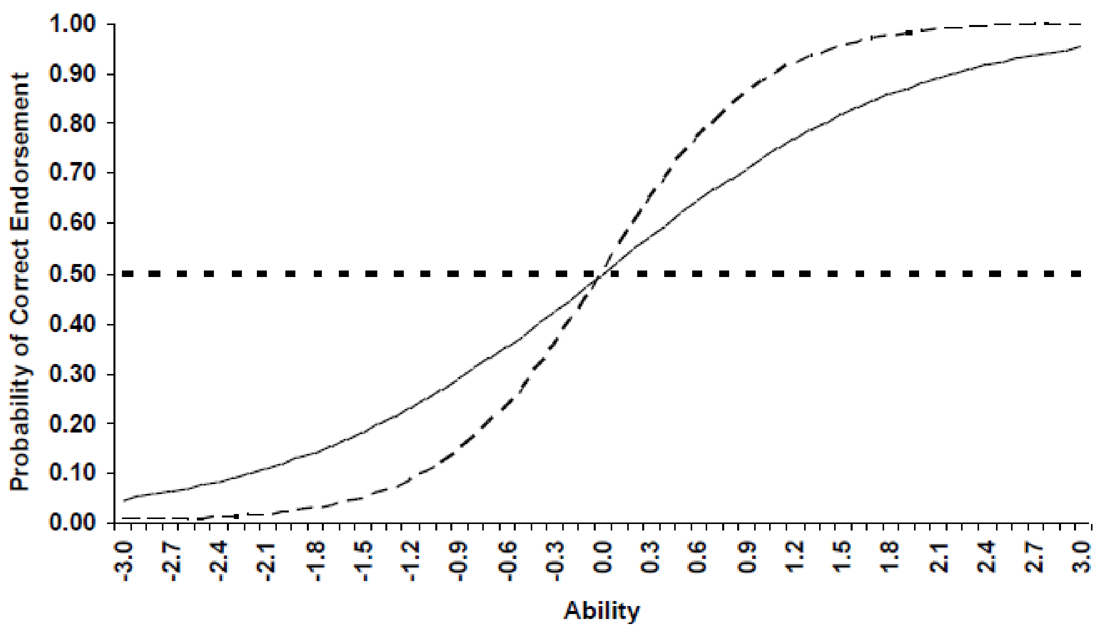


Figure 2. Sample Item Characteristic Curves with Varied Discriminations

Guidelines for Retaining Items

Several criteria were used to evaluate item validity. The first process was to identify items which demonstrated strong floor or ceiling effects in response rates >= 95%. Such items are not useful in

creating an item bank as there is little variability in whether students are successful on the item. In addition to evaluating the descriptive response rate, we estimated item-total correlations. Items with negative values are indicative of poor functioning such that it suggests individuals who correctly answer the question tend to have lower total scores. Similarly, items with low item-total correlations indicate the lack of a relation between item and total test performance. Items with correlations <.15 were flagged for removal. Following the descriptive analysis of item performance, difficulty and discrimination values from the IRT analyses were used to further identify items which were poorly functioning. Items were flagged for item revision if the item discrimination was negative or the item difficulty was greater than +4.0 or less than -4.0.

Secondary criteria were used in evaluating the retained items, which was comprised of a differential item function (DIF) analysis. DIF refers to instances where individuals from different groups with the same level of underlying ability significantly differ in their probability to correctly endorse an item. Unchecked, items included in a test which demonstrate DIF will produce biased test results. For the FAIR-FS assessments, DIF testing was conducted comparing: Black-White students, Latino-White students, Black-Latino students, students eligible for Free or Reduced Priced Lunch (FRL) with students not receiving FRL, and English Language Learner to non-English Language Learner students.

DIF testing was conducted with a multiple indicator multiple cause (MIMIC) analysis in Mplus (Muthén & Muthén, 2008); moreover, a series of four standardized and expected score effect size measures were generated using VisualDF software (Meade, 2010) to quantify various technical aspects of score differentiation between the gender groups. First, the signed item difference in the sample (SIDS) index was created, which describes the average unstandardized difference in expected scores between the groups. The second effect size calculated was the unsigned item difference in the sample (UIDS). This index can be utilized as supplementary to the SIDS. When the absolute value of the SIDS and UIDS values are equivalent, the differential functioning between groups is equivalent; however, when the absolute value of the UIDS is larger than SIDS, it provides evidence that the item characteristic curves for expected score differences cross, indicating that differences in the expected scores between groups change across the level of the latent ability score. The D-max index is reported as the maximum SIDS value in the sample, and may be interpreted as the greatest difference for any individual in the sample in the expected response. Lastly, an expected score standardized difference (ESSD) was generated, and was computed similar to a Cohen's (1988) *d* statistic. As such, it is interpreted as a measure of standard deviation difference between the groups for the expected score response with values of .2 regarded as small, .5 as medium, and .8 as large.

Linking Design & Item Response Analytic Framework

A common-item, non-equivalent groups design was used for collecting data in our pilot, calibration, and validation studies. A strength of this approach is that it allows for linking multiple test forms via common items. For each task, a minimum of twenty-percent of the total items within a form were identified as vertical linking items to create a vertical scale. These items served a dual purpose of not only linking forms across grades to each other, but also linking forms within grades to each other.

Because the tasks in the FAIR-FS were each designed for vertical equating and scaling we considered two primary frameworks for estimating the item parameters: 1) a multiple-group IRT of all test forms or 2) test characteristic curve equating. We chose the latter approach using Stocking and Lord (1983) to place the items on a common scale. All item analyses were conducted using Mplus software (Muthén & Muthén, 2008) with a 2pl independent items model. Because the samples used for data collection did not strictly adhere to the state distribution of demographics (i.e., percent limited English proficiency, Black, White, Latino, and eligible for free/reduced lunch), sample weights according to student demographics were used to inform the item and student parameter scores.

Norming Studies

Students from several districts throughout Florida participated in the common-item, non-equivalent groups linking study to estimate and evaluate the item parameters and student ability score distributions for each of the computer adaptive tasks (CAT) in the FAIR-FS. A total of 44,780 students in grades 3-12 across six districts in Florida participated in the calibration and validation studies which consisted of students taking the FAIR-FS tasks appropriate to levels of performance. Table 2 provides a breakdown of the sample sizes used by grade level for each of the FAIR-FS adaptive assessments. Average demographic information for the state in grades 3-10 was as follows: 41% White, 30% Hispanic, 23% Black, 6% Other; 60% eligible for free/reduced price lunch; 8% limited English proficient[2]. The sample demographics for our validation sample approximately reflected state demographics as it pertains to the percent of White, Black, and Hispanic students, percentage of English language learners (ELL) and percentage of students eligible for free/reduced price lunch (FRL). A particular nuance with assessment research is that the collected sample data may not precisely reflect the population of interest. To correct for observed imprecision in how well a sample reflects a population, sample weights are used to reduce bias and compensate for over- or under- representativeness of the sample. Subsequently, our analyses were informed by weights constructed by evaluating the proportion of individuals who existed across combinations of race/ethnicity, ELL status, and FRL status. This resulted in 16 unique weights applied to the data to account for the four levels of race/ethnicity (White, Black, Hispanic, Other), two levels of FRL status (eligible/not eligible), and two levels of ELL status (ELL/not ELL). In this way our analyses were able to more precisely reflect the distribution of Florida's demographics according to key demographic characteristics. Specific sample weight data used in this study are reported in Appendix A.

---

[2] Data sources: Race data from 2013-14 Survey 3, Florida Department of Education; Free/Reduced Lunch data from 2013-14 Survey 2 data, Florida Department of Education and Archive Data Core, Florida Center for Reading Research; English Language Learner data from Education Information and Accountability Services, Florida Department of Education and Archive Data Core, Florida Center for Reading Research.

Table 2

Sample Size by Grade for FAIR-FS Tasks

Grade	Vocabulary Knowledge	Word Recognition	Syntactic Knowledge	Reading Comprehension
3	502	651	962	2,723
4	570	586	857	2,679
5	519	697	981	2,721
6	606	652	865	3,835
7	599	612	617	3,683
8	597	613	616	3,814
9	813	1,054	1,053	3,964
10	574	1,109	869	3,787
Total	4,780	5,974	6,820	27,206

Score Definitions

Several different kinds of scores are provided in order to facilitate a diverse set of educational decisions. In this section, we describe the types of scores provided for each measure, define each score, and indicate its primary utility within the decision making framework of the FAIR-FS. An ability score and a percentile rank are provided for each task (WRT, VKT, RC, and SKT) at each time point. One probability of literacy success score is provided at each assessment period.

Probability of Literacy Success (PLS). The Probability of Literacy Success score indicates the likelihood that a student will reach end of year expectations in literacy. For the purposes of the FAIR-FS in the 2014-2015 school year, reaching expectations is defined as performing at or above the 40th percentile on the Stanford Achievement Test, Tenth Edition (SAT-10)[3]. The PLS is used to determine which students are at-risk for meeting grade level expectations by the end of the school year. In addition to providing a precise probability of reaching grade level outcomes, the PLS is color-coded:

[3] The FAIR-FS will be realigned after the 2014-2015 school year to the Florida Standards Assessment (FSA).

- red = the student is at high risk and needs supplemental and/or intensive instruction targeted to the student’s skill weaknesses

- yellow = the student may be at-risk and educators may consider differentiating instruction for the student and/or providing supplemental instruction

- green = the student is likely not at-risk and will continue to benefit from strong universal instruction

In the grades 3-12 FAIR-FS, the components that are included in the PLS are an aggregate of the individual student’s VKT, WRT, and RC scores.

Percentile Ranks. Percentile ranks can vary from 1 to 99, and they divide the distribution of scores from a large standardization sample (in this case a representative sample of students from Florida) into 100 groups that contain approximately the same number of observations in each group. Thus, a sixth grade student who scored at the 60th percentile would have obtained a score better than about 60% of the students in the standardization sample. The median percentile rank on all the tests of the grades 3-12 FAIR-FS is 50, which means that half the students in the standardization sample obtained a score above that point, and half scored below it. The percentile rank is an ordinal variable meaning that it cannot be added, subtracted, used to create a mean score, or in any other way mathematically manipulated. The median is always used to describe the midpoint of a distribution of percentile ranks. Since this score compares a student’s performance to other students within a grade level, it is meaningful in determining the skill strengths and skill weaknesses for a student as compared to other students’ performance.

Ability Scores. Each computer-adaptive task has an associated ability score. The ability score provides an estimate of a student’s development in a particular skill. This score is sensitive to changes in a student’s ability as skill levels increase or decrease. Ability scores in the grades 3-12 FAIR-FS span the development of each of four important skills: Word Recognition, Vocabulary Knowledge, Reading Comprehension, and Syntactic Knowledge. The range of the developmental scale for each task is 200 to 1000, with a mean of 500 and standard deviation of 100. This score has an equal interval scale that can be added, subtracted, and used to create a mean score. Therefore, this is the score that should be used to determine the degree of growth in a skill for individual students.

Reliability

Marginal Reliability

Reliability describes how consistent test scores will be across multiple administrations over time, as well as how well one form of the test relates to another. Because the FAIR-FS uses Item Response Theory (IRT) as its method of validation, reliability takes on a different meaning than from a Classical Test Theory (CTT) perspective. The biggest difference between the two approaches is the assumption made about the measurement error related to the test scores. CTT treats the error variance as being the same for all scores, whereas the IRT view is that the level of error is dependent on the ability of the individual. As such, reliability in IRT becomes more about the level of precision of measurement across ability, and it may sometimes be difficult to summarize the precision of scores in IRT with a single number. Although it is often more useful to graphically represent the standard error across ability levels to gauge the range of abilities for which the test is more or less informative, it is possible to estimate a generic estimate of reliability known as marginal reliability (Sireci, Thissen, & Wainer, 1991) with:

$$\bar{\rho} = \frac{\sigma_\theta^2 - \overline{\sigma_{e*}^2}}{\sigma_\theta^2}$$

where σ_θ^2 is the variance of ability score for the normative sample and $\overline{\sigma_{e*}^2}$ is the mean-squared error. Marginal reliability coefficients for the three FAIR-FS Screening tasks are reported in Table 3 by grade and assessment period.

Table 3

Marginal Reliability for FAIR-FS Screening Tasks of Vocabulary Knowledge, Word Recognition, and Reading Comprehension at the Fall, Winter, and Spring Administrations

	Vocabulary Knowledge			Word Recognition			Reading Comprehension		
Grade	Fall	Winter	Spring	Fall	Winter	Spring	Fall	Winter	Spring
3	.84	.86	.87	.73	.85	.89	.85	.86	.83
4	.81	.83	.86	.86	.84	.88	.76	.85	.89
5	.87	.87	.88	.87	.84	.90	.80	.83	.90
6	.85	.85	.86	.86	.85	.91	.84	.87	.91
7	.85	.85	.86	.86	.86	.91	.78	.83	.91
8	.83	.84	.84	.87	.83	.92	.81	.85	.92
9	.85	.82	.86	.88	.80	.91	.67	.78	.91
10	.85	.81	.84	.88	.78	.90	.76	.82	.92
All Grades	.91	.89	.90	.92	.88	.93	.86	.88	.93

Note. Reliability coefficients for the Fall and Winter Reading Comprehension scores are reflective of fixed item administrations. Spring reliability coefficients for Reading Comprehension are reflective of performance on the CAT version. Marginal reliability coefficients for Vocabulary and Word Recognition are reflective of CAT versions of the assessments.

Across all grades and assessment periods, the marginal reliability was quite high ranging from .86 for fall reading comprehension to .93 for spring word recognition and reading comprehension. Values of .80 are typically viewed as acceptable for research purposes while estimates at .90 or greater are acceptable for clinical decision making (Nunnally & Berstein, 1994). Marginal reliability coefficients for the diagnostic Syntactic Knowledge Task are reported in Table 4. Similar to the other tasks, marginal reliability coefficients were quite high across all grades ranging from .92 to .93.

Table 4

Syntactic Knowledge Marginal Reliability Coefficients

	Syntax		
Grade	Fall	Winter	Spring
3	.85	.87	.89
4	.88	.87	.88
5	.87	.88	.90
6	.88	.89	.91
7	.88	.89	.91
8	.91	.88	.92
9	.91	.87	.90
10	.91	.87	.90
All Grades	.93	.92	.93

Note. Reliability coefficients for all assessment periods are reflective of the CAT version of the assessment

Standard Error of Measurement

A standard error of measurement (SEM; Harvill, 2005) is an estimate that captures the amount of variance that might be observed in an individual student's performance if they were tested repeatedly. That is, on any particular day of testing, an examinee's score may fluctuate and only through repeated testing is it possible to get closer to one's true ability. Because it is not reasonable to test a student enough to capture his/her true ability, we can construct an interval by which we can observe the extent to which the score may fluctuate. The SEM is calculated with:

$$\text{SEM} = \sigma_x\sqrt{1-\rho^2}$$

where σ_x is the standard deviation associated with the mean for assessment *x*, and ρ^2 is the marginal reliability for the assessment. Means and SEM are reported in Tables 5-7 for the 3 Screening tasks, respectively.

Table 5

Means and Standard Error of Measurement for Vocabulary Knowledge Scores

		Fall		Winter		Spring	
Grade	N	Mean	SEM	Mean	SEM	Mean	SEM
3	466	380.28	29.30	393.07	27.98	413.82	25.91
4	486	431.77	28.42	439.80	28.63	453.59	26.85
5	423	469.14	29.17	473.85	28.12	482.07	26.89
6	639	492.40	29.23	498.09	29.17	505.10	27.05
7	632	521.95	29.24	518.13	29.34	529.92	26.97
8	681	550.11	29.60	540.88	30.88	551.98	29.40
9	1014	555.66	29.40	560.26	32.00	562.86	28.62
10	887	571.88	30.28	575.32	36.19	574.38	30.44

Table 6

Means and Standard Error of Measurement for Word Recognition Scores

		Fall		Winter		Spring	
Grade	N	Mean	SEM	Mean	SEM	Mean	SEM
3	470	341.36	29.72	351.25	29.79	377.59	24.21
4	491	407.69	31.06	405.81	30.43	427.49	29.73
5	426	437.77	30.92	440.94	30.42	466.91	27.06
6	646	465.32	31.28	458.53	31.06	490.20	26.41
7	634	498.42	32.22	482.32	31.74	518.74	27.85
8	690	531.50	32.88	515.55	36.63	555.32	27.06
9	1017	543.01	33.21	543.53	43.68	567.72	29.29
10	916	574.34	33.96	558.00	47.27	591.01	32.76

Table 7

Means and Standard Error of Measurement for Reading Comprehension Scores

		Spring	
Grade	N	Mean	SEM
3	325	386.03	28.69
4	322	440.07	32.96
5	302	497.25	36.49
6	431	499.96	37.63
7	426	524.45	39.67
8	461	571.71	48.61
9	703	583.06	39.26
10	626	589.72	44.65

Note. Data is only provided for Spring due to the CAT version only being administered in the Spring.

Means and standard error of measurement for the diagnostic Syntactic Knowledge Task are reported in Table 8.

Table 8

Means and Standard Error of Measurement for Syntactic Knowledge Scores

		Fall		Winter		Spring	
Grade	N	Mean	SEM	Mean	SEM	Mean	SEM
3	377	328.84	30.80	358.06	30.58	402.12	25.29
4	376	403.74	30.06	417.15	30.80	452.63	24.85
5	340	430.52	30.12	452.58	30.82	483.09	25.29
6	383	456.01	31.18	473.15	31.59	505.59	25.04
7	396	510.01	30.40	504.94	31.41	529.24	25.49
8	380	523.01	30.16	533.04	34.28	554.57	25.73
9	457	554.38	32.05	551.09	36.27	571.61	27.52
10	443	554.98	31.07	549.89	38.55	562.49	28.15

Test-Retest Reliability

The extent to which a sample of students performs consistently on the same assessment across multiple occasions is an indication of test-retest reliability. Reliability was estimated for students participating in the field testing of the FAIR-FS by correlating their ability scores across three assessments. Retest correlations for vocabulary and word recognition (Table 9) were the strongest between winter and spring while the fall-winter correlations were strongest for reading comprehension. Correlations between the fall and spring were the lowest, which is expected as a weaker correlation from the beginning of the year to the end suggests that students were differentially changing over time (i.e., lower ability students may have grown more over time compared to higher ability students). Retest correlations for the diagnostic Syntactic Knowledge Task are reported in Table 10. Similar to the Vocabulary Knowledge and Word Recognition Tasks, the strongest correlations between time-points were the winter-spring associations.

Table 9

FAIR-FS Screening Test-Retest Correlations for Vocabulary Knowledge, Word Recognition, and Reading Comprehension

	Vocabulary Knowledge			Word Recognition			Reading Comprehension		
Grade	Fall-Winter	Winter-Spring	Fall-Spring	Fall-Winter	Winter-Spring	Fall-Spring	Fall-Winter	Winter-Spring	Fall-Spring
3	.59	.61	.44	.46	.51	.31	.74	.66	.66
4	.58	.62	.51	.59	.62	.45	.83	.77	.71
5	.75	.74	.65	.63	.73	.64	.83	.77	.73
6	.60	.72	.51	.59	.65	.66	.85	.80	.77
7	.66	.69	.54	.65	.69	.73	.80	.79	.73
8	.63	.67	.63	.66	.72	.74	.81	.79	.71
9	.65	.64	.65	.65	.68	.76	.77	.72	.65
10	.62	.70	.64	.69	.70	.80	.75	.74	.66

Table 10

Test-Retest Correlations for Syntactic Knowledge Task

	Syntax		
Grade	Fall-Winter	Winter-Spring	Fall-Spring
3	.49	.55	.48
4	.62	.70	.56
5	.68	.75	.68
6	.63	.69	.65
7	.68	.74	.69
8	.66	.76	.70
9	.70	.73	.80
10	.67	.70	.72

Validity

Assessment of Model Fit

A first step in testing the validity of scores was to evaluate the dimensionality of item responses on each of the FAIR-FS tasks. An important assumption in IRT is unidimensionality, which states that a score from a test can only have meaning if the items measure one dimension. Connected to this assumption is the framework of local item independence, which requires that, for a given level of individual ability, individual responses to a set of items are statistically independent of each other (Hattie, Krakowski, Rogers, & Swaminathan, 1996). McDonald (1979) suggested that a weaker principle of independence should be used, whereby only the covariances must be zero, and that the relationship between moments did not need to be considered. Stout (1990) extended the logic of weak local independence to argue for “essential unidimensionality” rather than ascribing to more stringent standards. Conceptually, Stout argued that a test is unidimensional if, for a given level of ability, the average covariance over pairs of items on the test is small in magnitude, as opposed to zero. Essential unidimensionality may be formally assessed through a variety of methods including parametric and non-parametric exploratory and confirmatory factor analysis. For the FAIR-FS tasks, a parametric confirmatory factor analysis was run on scores for different forms of each task by grade level. Because a planned missing data design was used, the covariance coverage was necessarily low. A planned missing data design with a large number of items frequently precludes a factor analysis of the full item response matrix when using the weighted least squares multivariate estimator. This estimator is necessary to produce commonly used fit indices for confirmatory factor analysis. Subsequently, the factor analysis was carried out by form and grade within each task. The comparative fit index (CFI), Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA) were used to evaluate model fit for the Vocabulary Knowledge, Word Recognition, and Syntax Knowledge tasks. CFI and TLI values of at least .90 are considered acceptable as are RMSEA values less than .10. For the Reading Comprehension task, we tested the extent to which a unidimensional model fit better than a testlet model. The two models were compared using the AIC and BIC indices.

Fit statistics for Vocabulary Knowledge, Word Recognition, and Syntax Knowledge are reported in Tables 11, 12, and 13, respectively. Results demonstrate that item responses across forms and grades converge on an essentially unidimensional construct for the three tasks.

Table 11

Fit statistics by form and grade for the Vocabulary Knowledge Task

Grade	Form	χ^2	df	*p*-value	RMSEA	RMSEA LB	RMSEA UB	RMSEA *p*-value	CFI	TLI
3	A	202.51	170	0.045	0.020	0.000	0.032	1.00	0.96	0.96
	B	175.65	152	0.092	0.019	0.000	0.031	1.00	0.97	0.96
4	A	195.50	189	0.358	0.009	0.000	0.022	1.00	0.99	0.99
	B	214.65	189	0.097	0.017	0.000	0.027	1.00	0.97	0.97
5	A	199.62	189	0.284	0.011	0.000	0.024	1.00	0.98	0.98
	B	169.92	170	0.487	0.000	0.000	0.022	1.00	1.00	1.00
6	A	385.84	377	0.366	0.006	0.000	0.016	1.00	0.99	0.99
	B	441.40	377	0.012	0.017	0.008	0.023	1.00	0.96	0.96
7	A	207.17	189	0.174	0.014	0.000	0.025	1.00	0.95	0.94
	B	219.36	189	0.064	0.018	0.000	0.028	1.00	0.98	0.98
8	A	216.55	189	0.083	0.017	0.000	0.027	1.00	0.97	0.97
	B	228.64	189	0.026	0.021	0.008	0.029	1.00	0.94	0.93
9	A	215.70	189	0.089	0.014	0.000	0.023	1.00	0.98	0.98
	B	225.72	189	0.035	0.017	0.005	0.002	1.00	0.96	0.96
10	A	204.25	189	0.212	0.012	0.000	0.022	1.00	0.98	0.98
	B	232.27	170	0.001	0.028	0.018	0.037	1.00	0.89	0.88

Note. df = degrees of freedom; RMSEA = root mean square error of approximation; LB = lower bound; UB = upper bound; CFI = comparative fit index; TLI = Tucker-Lewis index.

Table 12

Fit statistics by grade and form for the Word Recognition Task

Grade	Form	χ^2	df	*p*-value	RMSEA	RMSEA UB	RMSEA LB	RMSEA *p*-value	CFI	TLI
3	A	233.54	152	0.000	0.042	0.031	0.052	0.91	0.93	0.92
	B	130.20	104	0.042	0.027	0.006	0.041	1.00	0.96	0.95
4	A	99.27	65	0.004	0.044	0.025	0.061	0.71	0.90	0.87
	B	135.26	119	0.146	0.021	0.000	0.036	1.00	0.95	0.94
5	A	173.02	152	0.117	0.020	0.000	0.030	1.00	0.96	0.95
	B	81.14	65	0.085	0.027	0.000	0.044	0.99	0.94	0.93
6	A	478.14	377	0.000	0.020	0.014	0.026	1.00	0.93	0.93
	B	425.31	350	0.004	0.018	0.011	0.024	1.00	0.94	0.94
7	A	189.75	152	0.020	0.029	0.012	0.041	1.00	0.90	0.89
	B	86.31	90	0.590	0.000	0.000	0.028	1.00	1.00	1.00
8	A	179.94	152	0.060	0.025	0.000	0.038	1.00	0.91	0.90
	B	154.74	135	0.118	0.022	0.000	0.036	1.00	0.95	0.94
9	A	198.25	152	0.007	0.024	0.013	0.032	1.00	0.96	0.95
	B	140.16	152	0.745	0.000	0.000	0.016	1.00	1.00	1.00
10	A	196.33	152	0.009	0.025	0.013	0.034	1.00	0.92	0.91
	B	102.48	77	0.028	0.029	0.010	0.040	1.00	0.88	0.86
	C	404.31	377	0.159	0.017	0.000	0.029	1.00	0.95	0.94

Note. df = degrees of freedom; RMSEA = root mean square error of approximation; LB = lower bound; UB = upper bound; CFI = comparative fit index; TLI = Tucker-Lewis index.

Table 13

Fit statistics by grade and form for the Syntax Knowledge Task

Grade	Form	χ^2	df	*p*-value	RMSEA	RMSEA UB	RMSEA LB	RMSEA *p*-value	CFI	TLI
3	A	189.18	170	0.149	0.011	0.000	0.019	1.00	0.94	0.93
	B	198.78	152	0.007	0.018	0.010	0.024	1.00	0.96	0.96
4	A	188.69	135	0.001	0.022	0.014	0.029	1.00	0.90	0.88
	B	167.71	152	0.182	0.011	0.000	0.020	1.00	0.97	0.97
5	A	211.22	170	0.017	0.016	0.007	0.022	1.00	0.92	0.91
	B	177.81	152	0.075	0.013	0.000	0.021	1.00	0.97	0.96
6	A	205.98	170	0.031	0.160	0.005	0.023	1.00	0.96	0.95
	B	293.34	230	0.003	0.018	0.011	0.024	1.00	0.95	0.94
	C	231.39	170	0.001	0.020	0.013	0.027	1.00	0.93	0.93
7	A	160.33	170	0.691	0.000	0.000	0.015	1.00	1.00	1.00
	B	176.75	170	0.345	0.008	0.000	0.020	1.00	0.98	0.97
8	A	304.36	170	0.000	0.036	0.029	0.042	1.00	0.82	0.80
	B	275.77	135	0.000	0.041	0.034	0.048	0.98	0.77	0.74
9	A	184.00	170	0.219	0.009	0.000	0.017	1.00	0.99	0.99
	B	221.00	170	0.005	0.017	0.010	0.023	1.00	0.92	0.91
10	A	199.47	170	0.061	0.014	0.000	0.022	1.00	0.93	0.93
	B	160.32	135	0.068	0.015	0.000	0.023	1.00	0.88	0.86

Note. df = degrees of freedom; RMSEA = root mean square error of approximation; LB = lower bound; UB = upper bound; CFI = comparative fit index; TLI = Tucker-Lewis index.

Model fit comparisons between the unidimensional and testlet models for the Reading Comprehension Task are reported in Table 14.

Table 14

AIC and BIC values for the unidimensional and testlet models in Reading Comprehension by grade

Grade	Model	AIC	BIC	adjusted-BIC
3	Unidimensional	103845	106019	104851
	Testlet	103672	106928	105177
4	Unidimensional	113842	115987	114830
	Testlet	113553	116765	115033
5	Unidimensional	101720	130349	102539
	Testlet	101471	104130	102700
6	Unidimensional	151414	153927	152649
	Testlet	150809	154579	152663
7	Unidimensional	121206	123155	122158
	Testlet	-	-	-
8	Unidimensional	141907	144093	142981
	Testlet	141541	144820	143153
9	Unidimensional	143848	146261	145041
	Testlet	143673	147293	145463
10	Unidimensional	122108	124454	123259
	Testlet	121811	125330	123538

Note. Grade 7 Testlet model did not converge.

Results from this comparison based on AIC and BIC were mixed. The AIC suggests that the testlet model should be used while the BIC and adjusted BIC values were smaller for the unidimensional model. Although the indices provide mixed information, the penalty term is greater in the BIC compared to the AIC. Due to the penalty difference, the BIC is a more conservative estimate and given the results above it was deemed more appropriate for model selection. Subsequently, the unidimensional model was retained.

Criterion Validity

Criterion validity describes how well scores on one assessment relate to other theoretically relevant constructs, both concurrently and predictively. Concurrent validity was evaluated by correlating scores from the tasks amongst each other while predictive validity was evaluated by using the FAIR-FS tasks to predict later reading comprehension performance on the SAT-10.

Concurrent Validity

Reading and language skills tend to have moderate associations between them; thus, the expectation of the FAIR-FS Vocabulary Knowledge, Word Recognition, and Syntactic Knowledge Tasks would be that stronger associations with reading comprehension would be observed compared to more moderate associations with each other. Correlation results are reported in Table 15.

Table 15

Bivariate Associations among FAIR-FS Tasks

Grade	Measure	Reading Comprehension	Vocabulary	Word Recognition	Syntax
3	Reading Comprehension	1.00			
	Vocabulary Knowledge	.60	1.00		
	Word Recognition	.42	.37	1.00	
	Syntax Knowledge	.48	.38	.30	1.00
4	Reading Comprehension	1.00			
	Vocabulary Knowledge	.42	1.00		
	Word Recognition	.43	.30	1.00	
	Syntax Knowledge	.52	.35	.29	1.00
5	Reading Comprehension	1.00			
	Vocabulary Knowledge	.58	1.00		
	Word Recognition	.40	.37	1.00	
	Syntax Knowledge	.57	.44	.31	1.00
6	Reading Comprehension	1.00			
	Vocabulary Knowledge	.54	1.00		
	Word Recognition	.48	.36	1.00	
	Syntax Knowledge	.58	.45	.36	1.00

7	Reading Comprehension	1.00			
	Vocabulary Knowledge	.46	1.00		
	Word Recognition	.45	.38	1.00	
	Syntax Knowledge	.60	.44	.42	1.00
8	Reading Comprehension	1.00			
	Vocabulary Knowledge	.49	1.00		
	Word Recognition	.49	.40	1.00	
	Syntax Knowledge	.59	.44	.46	1.00
9	Reading Comprehension	1.00			
	Vocabulary Knowledge	.53	1.00		
	Word Recognition	.55	.53	1.00	
	Syntax Knowledge	.63	.58	.54	1.00
10	Reading Comprehension	1.00			
	Vocabulary Knowledge	.50	1.00		
	Word Recognition	.49	.51	1.00	
	Syntax Knowledge	.59	.55	.57	1.00

Predictive Validity

The predictive validity of the Screening tasks to the SAT-10 Reading Comprehension test for grades 3-12 was addressed through a series of linear and logistic regressions. The linear regressions were run two ways. First, a correlation analysis was used to evaluate the strength of relations between each of the Screening tasks' ability scores with the SAT-10. Second, a multiple regression was run to estimate the total amount of variance that the linear combination of the predictors explained in SAT-10 reading comprehension performance. Results from the linear regression analyses are reported in Table 16.

Table 16

Bivariate Correlations between FAIR-FS Screening Tasks and SAT-10. Percent Variance Explained in SAT-10 by FAIR-FS Vocabulary, Word Recognition, and Reading Comprehension

Grade	Vocabulary Knowledge	Word Recognition	Reading Comprehension	Total R^2
3	.56	.43	.74	.62
4	.45	.39	.71	.56
5	.57	.41	.74	.59
6	.53	.46	.71	.53
7	.43	.43	.66	.45
8	.46	.47	.67	.48
9	.51	.55	.60	.47
10	.47	.51	.57	.39

For the logistic regressions, students' performance on the SAT-10 Reading Comprehension test was coded as '1' for performance at or above the 40th percentile, and '0' for scores below this target. This dichotomous variable was then regressed on a combination of vocabulary knowledge, word recognition, and reading comprehension scores at each grade level. Further, we evaluated the classification accuracy of scores from the FAIR-FS as it pertains to risk status on the SAT-10. By dichotomizing the combination of screening task scores as '1' for not at-risk for reading difficulties and '0' for at-risk for reading difficulties, students could be classified based on their dichotomized performances on both. As such, students could be identified as not at-risk on the combination of screening tasks and demonstrating grade level performance on the SAT-10 (i.e., specificity or true-negatives), at-risk on the combination of screening task scores and below grade level performance on the SAT-10 (i.e., sensitivity or true-positives), not at-risk based on the combination of screening task scores and not at grade level on the SAT-10 (i.e., false negative error), or at-risk on the combination of screening task scores and at grade level on the SAT-10 (i.e., false positive error). Classification of students in these categories allows for the evaluation of cut-points on the combination of screening tasks (i.e., PLS) to determine which PLS cut-point maximizes predictive power

The concept of risk can be viewed in many ways, including the concept as a "percent chance" which is a number between 0 and 100, with 0 meaning there is no chance that a student will develop a problem,

and 100 being there is no chance the student will not develop a problem. When attempting to identify children who are “at-risk” for poor performance on some type of future measure of reading achievement, this is typically a yes/no decision based upon a “cut-point” along a continuum of risk. Oftentimes this future measure of achievement is a state’s high-stakes assessment, which typically provides a standard score that describes the performance of each student. Grade-level cut-points are chosen that determine whether a student has passed or failed the state-wide assessment.

Decisions concerning appropriate cut-points for screening measures are made based on the level of correct classification that is desired from the screening assessments. While a variety of statistics may be used to guide such choices (e.g., sensitivity, specificity, positive and negative predictive power; see Schatschneider, Petscher, & Williams, 2008), negative predictive power was utilized to develop the FAIR-FS cut-points. Negative predictive power is the percentage of students who are identified as “not at-risk” on the screening assessments that end up not passing based the outcome assessment. Predictive power is not considered to be a property of the screening assessments since it is known to fluctuate given the proportion of individuals who are at-risk on the selected outcome (Streiner, 2003).

The cut-point selected for the grades 3-12 FAIR-2009 (used in the State of Florida from 2009-2014, Florida Department of Education, 2009) was negative predictive power of 0.85, meaning that at least 85% of students identified as “not at-risk” on the FAIR-2009 (i.e., FSP >= 0.85) would achieve at least a Level 3 on the Florida Comprehensive Assessment Test (FCAT) reading assessment at the end of the year. Greater emphasis was placed on negative predictive power than positive predictive power because the consequences of being identified as “at-risk” when the student is not actually at-risk are so much less than identifying students as “not at-risk” when they are actually at-risk for below grade-level performance on the FCAT. Prior research (Foorman & Petscher, 2010a; Foorman & Petscher, 2010b; Petscher & Foorman, 2011) demonstrated the technical adequacy of using .85 as an appropriate cut-point for risk on the FAIR 2009. As part of a continuing evaluation of the classification accuracy of FAIR 2009 scores, Petscher and Foorman (2011) found that an alternative cut-point (i.e., .70) could be used to maintain high negative predictive power and also minimize identification errors. As it pertains to the FAIR-FS, we tested the extent to which using a .85 cut-point for a student being identified as not at-risk yielded a negative predictive power value of at least 85%. Similarly, we also tested (a) how high negative predictive power would be estimated when using a cut-point of .70, and (b) whether identification errors could be reduced. A summary of the classification results for FAIR-FS are reported in Table 17.

Table 17

Classification Accuracy of the Probability of Literacy Success (PLS) in Grades 3-12 using .85 and .70 Cut-Points

Cut-Point	Grade	SE	SP	PPP	NPP	OCC	Base Rate
.85	3	.95	.54	.59	.94	.71	.41
	4	.95	.58	.52	.96	.70	.32
	5	.94	.60	.56	.95	.72	.35
	6	.96	.39	.61	.91	.68	.50
	7	.98	.46	.55	.97	.67	.40
	8	.94	.46	.54	.92	.64	.39
	9	.93	.50	.38	.96	.61	.25
	10	.87	.52	.42	.91	.62	.28
.70	3	.85	.69	.66	.87	.76	.41
	4	.77	.74	.59	.88	.75	.32
	5	.83	.76	.65	.89	.78	.35
	6	.92	.56	.68	.87	.86	.50
	7	.91	.60	.61	.91	.73	.40
	8	.85	.67	.62	.88	.74	.39
	9	.76	.69	.45	.90	.71	.25
	10	.64	.74	.49	.84	.71	.28

Note. SE= Sensitivity, SP = Specificity, PPP = Positive Predictive Power, NPP = Negative Predictive Power, OCC = Overall Correct Classification. Students in Grades 11 and 12 are classified according to Grade 10 criteria.

Note that when using either the .85 or .70 cut-points the negative predictive power is above .85; yet, when the .85 cut-point is used, the specificity and positive predictive power are relatively low. The consequence of a low specificity value is that many students are required to take one or more additional tasks; in the present sample this would result in between 40% and 61% of students identified as false positives and required to take the Diagnostic tasks. Conversely, if a .70 cut-point is used, this error rate range reduces from 40%-61% down to 24% -44%. Coupled with a false positive reduction is an increase

in the positive predictive power and the overall correct classification. Although there is some loss of precision in the sensitivity, the negative predictive power maintains a high value to ensure that students who are identified as not at-risk have a high likelihood of being successful on end of year outcomes (i.e., 40th percentile or greater on the SAT-10).

Contextual Considerations in the Probability of Literacy Success (PLS)

The PLS score is a useful indicator of evaluating an individual student's likelihood of meeting a pre-set expectation on a selected outcome. When using the FAIR-FS PLS for the early identification of reading difficulties, it is important for the user to be aware of two key considerations: 1) how "meeting expectations" is defined on the selected outcome, and 2) what the impact is of "meeting expectations" on the distribution (i.e., the mean and spread of scores) of the PLS.

Defining "meeting expectations". As noted in the previous section on predictive validity, scores on the FAIR-FS are used to estimate the probability a student will perform at or above the 40th percentile on the SAT-10 Reading Comprehension outcome. The decision to use the 40th percentile to define "meeting expectations" is two-fold. First, the 40th percentile was used in the original version of FAIR (Florida Department of Education, 2009-2014) to define success on the SAT-10 Reading Comprehension. Second, the 40th percentile was used by several states as the criterion for student performance success during Reading First (Petscher, Kim, Foorman, 2011). Subsequently, this threshold was adopted for the purposes of screening for reading difficulties in the FAIR-FS to maintain consistency with previous standards. It is important to recognize that while the 40th percentile is a reasonable standard for defining expectations of success, it is also possible to change the standard. The choice of how to define expectations in a universal screener should be based on a confluence of substantive theory, measurement theory, and policy. By defining "meeting expectations" as performance at or above the 40th percentile, it is expected that in a sample of students who are normally distributed in their reading comprehension skills, approximately 60% of students should perform at or above the 40th percentile. The implication of the 40th percentile in a sample of students with normally distributed reading skills is that most students would be considered to be "meeting expectations." Should the operational definition on the outcome change, the percent of individuals who are "meeting expectations" will also change. Suppose that the 70th percentile is used as the target on the SAT-10 rather than the 40th percentile. In this instance, we would only expect 30% of students in a sample with normally distributed reading skills to "meet expectations" compared to 60% when using the 40th percentile. This simple example highlights the fact that while the qualitative designation of "meeting expectations" is the same across conditions, the number of students actually achieving at or above that pre-set level will vary depending on the pre-set level.

Understanding the Impact of "meeting expectations" on PLS. A change in definition of "meeting expectations" will influence the number of students who meet the defined threshold. Therefore, it is important to understand how such changes may influence the PLS. Returning to the conceptual example presented above, where "meeting expectations" on the SAT-10 could vary from

performing at either the 40th or 70th percentiles, questions educators might ask are: “What impact does varying criteria have on the distribution of the PLS?” “What distribution of the PLS *should* we expect?”

To answer these questions we refer to data from our norming study presented in the section on predictive validity. From the norming sample, we reproduced the logistic regressions which created the PLS and display the distribution of PLS scores for students in grades 3-10.

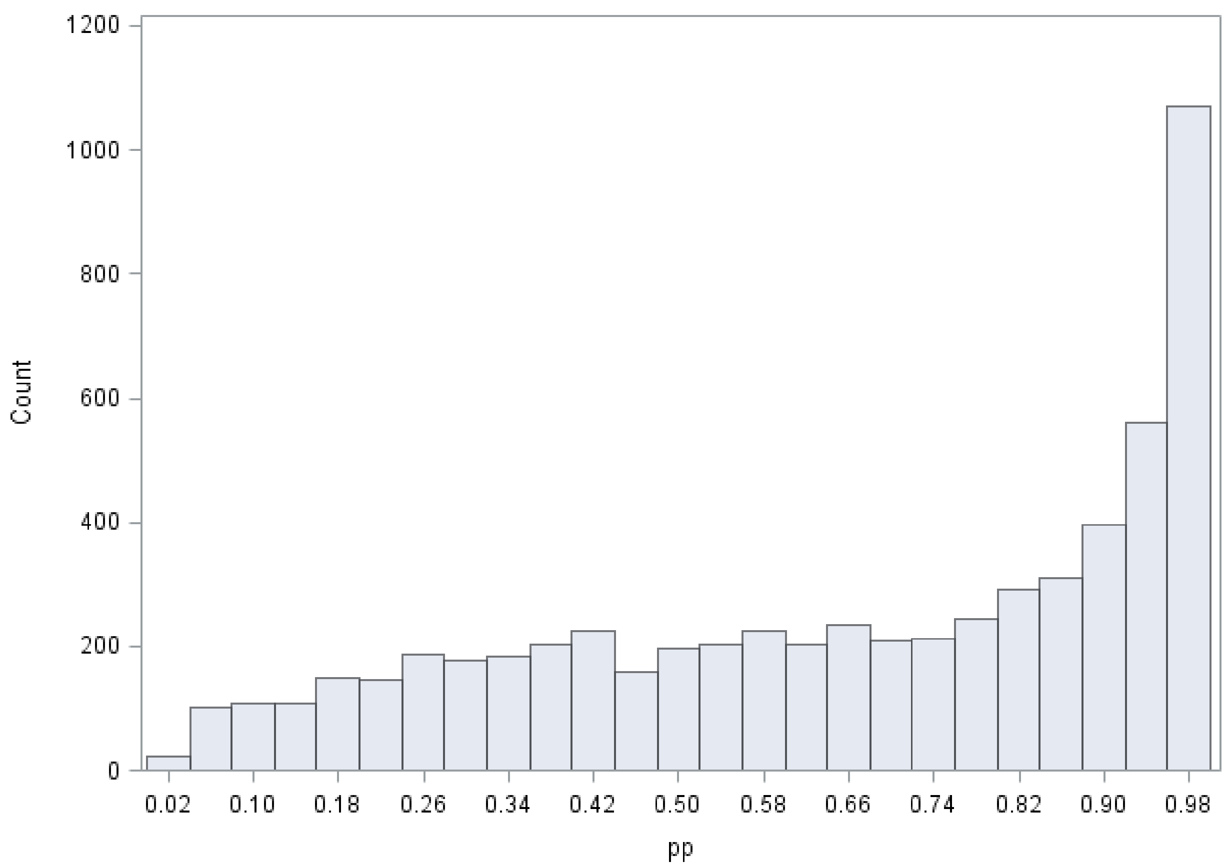


Figure 3. Distribution of PLS (pp) scores for students in grades 3-10 based on SAT-10 40th percentile

Note that Figure 3 shows that the PLS scores, which are predicted probabilities on the x-axis, are not normally distributed. The average PLS score for all students in the norming sample was .66 indicating that on average, students had a 66% chance of performing at or above the 40th percentile on the SAT-10. If we change the criteria of meeting expectations of the SAT-10 to the 70th percentile, we see a different distribution of PLS scores for students across grades 3-10.

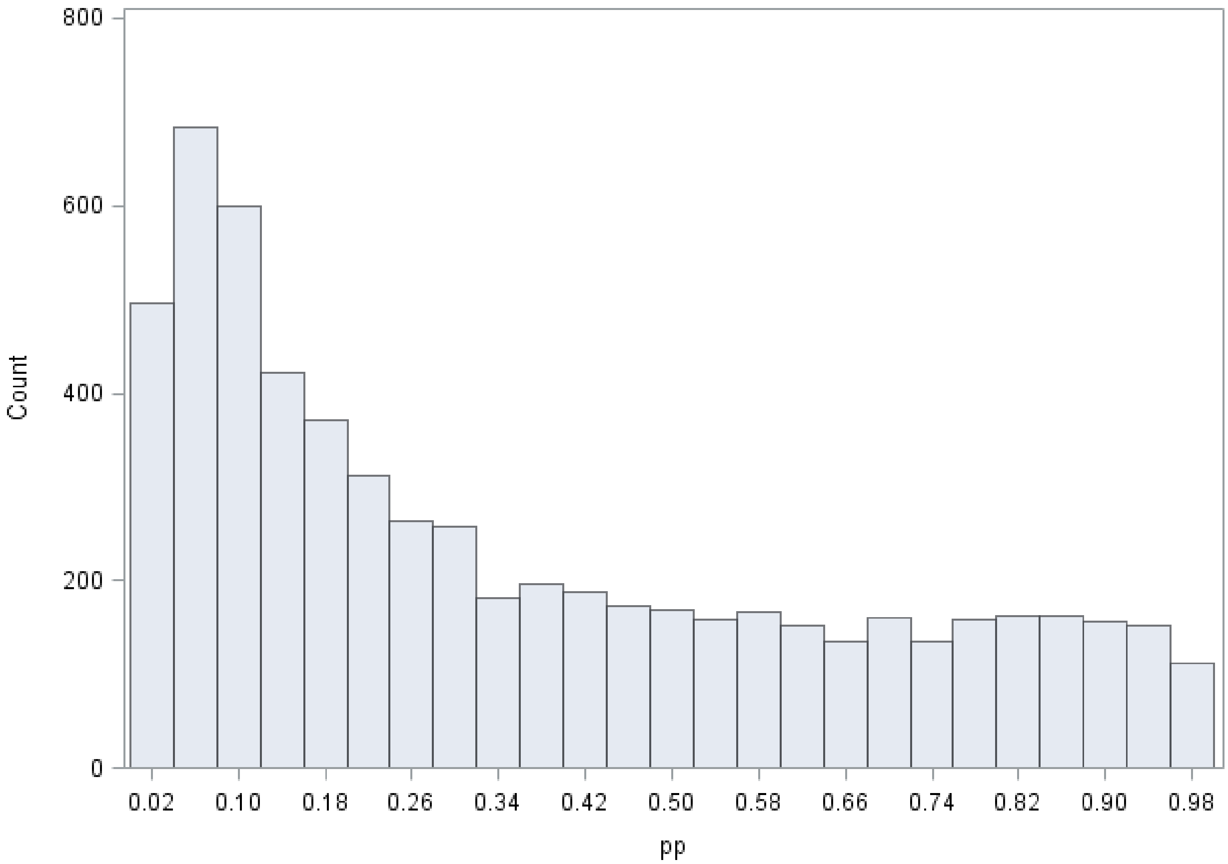


Figure 4. Distribution of PLS (pp) scores for students in grades 3-10 based on SAT-10 70th percentile

Notice that the distributions of PLS when the target for "meets expectations" on the SAT-10 changes from the 40th percentile in Figure 3 to the 70th percentile in Figure 4 results in more scores at the lower end of the distribution. The mean PLS for all students in Figure Y is .36, indicating that on average students had a 36% chance of performing at or above the 70th percentile on the SAT-10. Taken together, Figures 3 and 4 demonstrate that changing the target changes both the average likelihood of "meeting expectations" as well as the distribution of the PLS. When broken out by grade level (Appendix B and C), the same phenomenon exists whereby more students have higher PLS scores for the 40th percentile (Appendix B) target compared to the 70th percentile target (Appendix C).

The related question of, "What distribution of PLS should we expect?" is answered by statistical theory. Results from a logistic regression analysis do not automatically generate a probability value. Rather, logistic regression analysis from commonly used statistical software packages produces a log odds value for the model coefficients. These log odds are subsequently converted to a probability value using:

$$pp = \frac{e^{\ln(\widehat{OR})}}{1 + e^{\ln(\widehat{OR})}}$$

This equation states that a probability is calculated as a function of Euler's constant (i.e., *e;* 2.718) applied to the log odds. It is important to recognize that as it pertains to log odds and probabilities, logistic regression only makes distributional assumptions about the log odds. That is, *log odds are assumed to be normally distributed in logistic regression but predicted probability values are not*. To demonstrate, consider again the PLS scores for the norming sample which appear in Figures 3 and 4. In Figures 5 and 6 below, we place the distribution of the log odds above the distribution of the PLS.

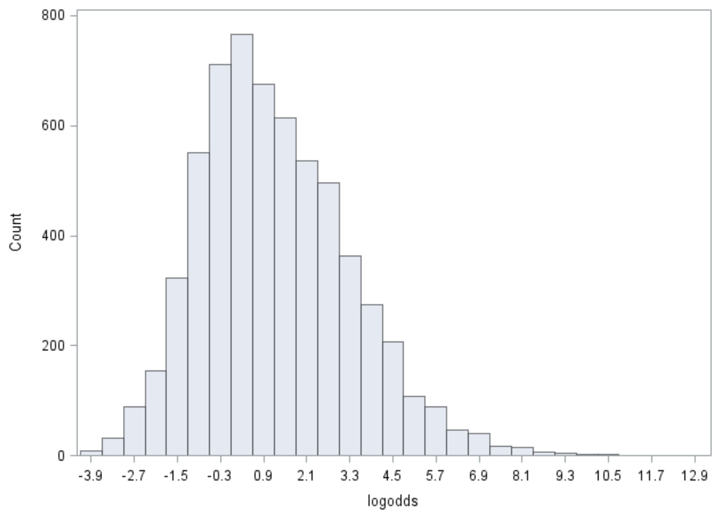


Figure 5. Distribution of log odds for the 40th percentile of SAT-10

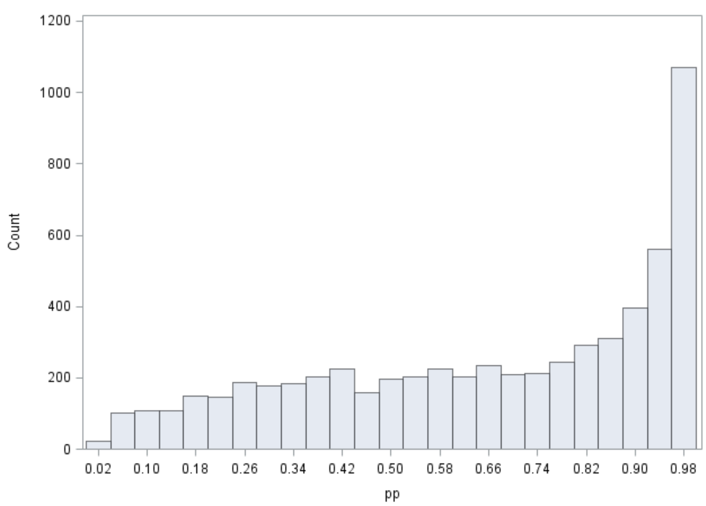


Figure 6. Distribution of PLS (pp) scores for the 40th percentile of SAT-10

Notice how the distribution of the log odds in Figure 5 is normally distributed but the PLS is not. The skewness value of -.51 and kurtosis of .59 are both small, indicating that the distribution is, indeed,

approximately normally distributed. If we compare the distribution of the log odds and PLS using the 70th percentile on the SAT-10 (Figures 7 and 8) we see a relatively similar phenomenon where the log odds are normally distributed and the PLS are not.

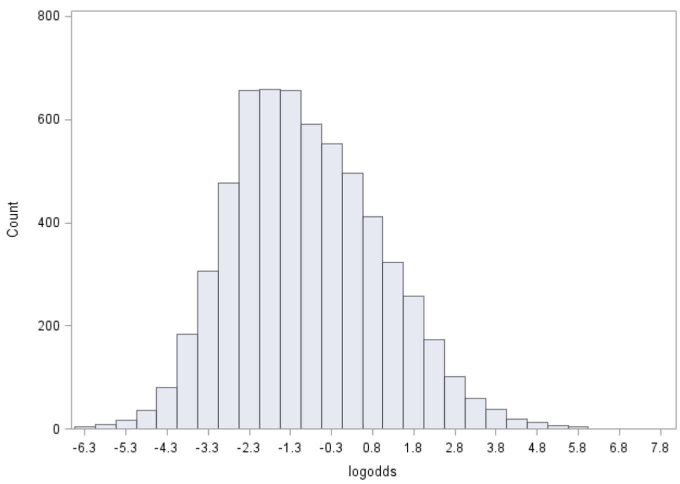


Figure 7. Distribution of log odds for the 70th percentile of SAT-10

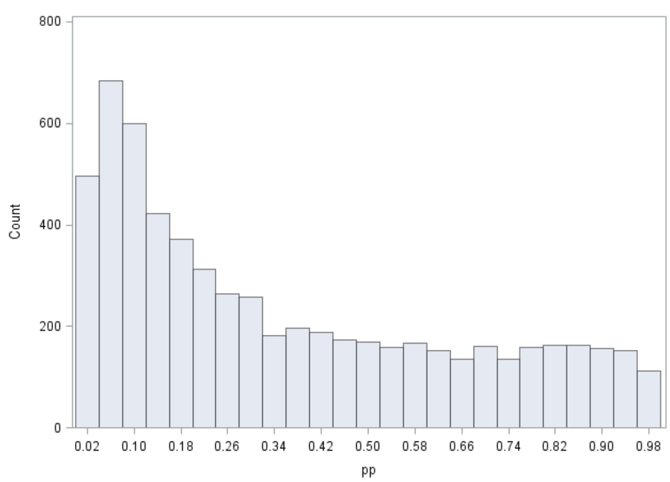


Figure 8. Distribution of PLS (pp) for the 70th percentile of SAT-10

The difference between log odds in Figure 7 and 8 can be seen in the mean scores; the mean log odds in Figure 7 is 1.23 compared to -.85 in Figure 8. This difference in mean values is instructive. Although both sets of log odds are normally distributed, the mean for one is much higher than the other. If the "meeting expectations" target is set at the 40th percentile, students have a higher, average log odds (i.e., 1.23) compared to if the target is set at the 70th percentile (i.e., -.85). Such a finding is expected! When higher standards are set for "meeting expectations" the result is that fewer individuals are able to meet

that target, thus a lower average log odds and PLS. The reverse is true when the standard is lower. The contextual considerations for using PLS illustrated in this section highlight that it is important to keep in mind the target being used for “meeting expectations”. For the FAIR-FS, the current threshold is the 40th percentile on the SAT-10. As noted previously, this was used to be in line with the previous version of the FAIR as well as with the standard used by the state of Florida in defining success during the *Reading First* initiative. As states move toward more demanding standards, the definition of “meeting expectations” may shift to the 50th percentile on a norm-referenced test.

Differential Accuracy of Prediction

An additional component of checking the validity of cut-points and scores on the assessments involved testing differential accuracy of the regression equations across different demographic groups. This procedure involved a series of logistic regressions predicting success on the SAT-10 test (i.e., at or above the 40th percentile). The independent variables included a variable that represented whether students were identified as not at-risk (PLS ≥ .70; coded as ‘1’) or at-risk (PLS < .70; coded as ‘0’) on the combination of screening task scores, a variable that represented a selected demographic group, as well as an interaction term between the two variables. A statistically significant interaction term would suggest that differential accuracy in predicting end-of-year performance existed for different groups of individuals based on the risk status determined by the screening assessment. For the combination of FAIR-FS screening task scores, differential accuracy was separately tested for Black and Latino students as well as for students identified as English Language Learners (ELL) and students who were eligible for Free/Reduced Price Lunch (FRL).

When testing for differential accuracy between Black and White students (Table 18), a significant effect for the interaction between the PLS cut-point and minority status existed in grade 4 (p = .003). This finding indicated that for the sample tested at the winter assessment period, White students with a PLS above the cut-point had a 92% chance of being at or above the 40th percentile on the SAT-10 compared to Black students above the cut-point on the PLS who had a 76% chance of being at or above the 40th percentile on the SAT-10. This translates into a 16% advantage in success for White students in grade 4, but we should note that replication will be needed across multiple administrations with a larger sample to evaluate the extent to which this phenomenon continues to exist.

When testing for differential accuracy between Hispanic and White students (Table 19), a significant effect for the interaction between the PLS cut-point and minority status existed in grades 8 and 10 (p = .015 and .02, respectively). This finding indicated that for the sample tested at the winter, White students in grade 8 with a PLS above the cut-point had an 87% chance of being at or above the 40th percentile on the SAT-10 compared to Hispanic students above the cut-point on the PLS who had an 89% chance of being at or above the 40th percentile on the SAT-10. This translates into a 3% advantage in success for Hispanic students in grade 8. Similarly, White students in grade 10 with a PLS above the cut-point had an 82% chance of being at or above the 40th percentile on the SAT-10 compared to Hispanic students above the cut-point on the PLS who had an 86% chance of being at or above the 40th percentile on the SAT-10. This translates into a 4% advantage in success for Hispanic students in grade

10. The findings from these two grades should be interpreted with caution as the mean difference in expected probability scores is quite small; thus, replication will be needed across multiple administrations with a larger sample to evaluate the extent to which this phenomenon continues to exist.

Table 18

Differential Accuracy for FAIR-FS Screening Tasks by Grade: Black-White (BW)

Grade	Parameter	df	Estimate	SE	χ^2	*p*-value
3	Intercept	1	-0.33	0.28	1.39	0.239
	PLS	1	3.69	0.65	32.12	<.001
	BW	1	-0.19	0.34	0.32	0.573
	PLS *BW	1	-1.31	0.77	2.86	0.091
4	Intercept	1	-0.66	0.33	3.98	0.046
	PLS	1	3.05	0.48	41.02	<.001
	BW	1	0.53	0.40	1.70	0.192
	PLS *BW	1	-1.78	0.60	8.83	0.003
5	Intercept	1	-0.31	0.27	1.37	0.243
	PLS	1	3.06	0.48	40.88	<.001
	BW	1	-0.33	0.34	0.91	0.340
	PLS *BW	1	-0.64	0.61	1.09	0.296
6	Intercept	1	-0.41	0.17	6.01	0.014
	PLS	1	2.62	0.34	59.29	<.001
	BW	1	-0.48	0.26	3.34	0.068
	PLS *BW	1	-0.85	0.57	2.22	0.137
7	Intercept	1	-0.31	0.18	2.98	0.085
	PLS	1	3.10	0.44	48.81	<.001

	BW	1	-0.14	0.28	0.25	0.615
	PLS *BW	1	-0.94	0.62	2.28	0.131
8	Intercept	1	-0.10	0.17	0.34	0.563
	PLS	1	1.97	0.29	46.72	<.001
	BW	1	-0.39	0.26	2.21	0.137
	PLS *BW	1	-0.09	0.49	0.04	0.849
9	Intercept	1	0.28	0.22	1.62	0.203
	PLS	1	2.31	0.42	30.23	<.001
	BW	1	-0.25	0.33	0.59	0.442
	PLS *BW	1	-0.38	0.59	0.42	0.517
10	Intercept	1	0.55	0.23	5.48	0.019
	PLS	1	0.99	0.30	11.05	0.001
	BW	1	-0.71	0.32	4.90	0.027
	PLS *BW	1	0.53	0.44	1.43	0.233

Note. PLS cut-off is .70. Estimates based on .85 cut-off approximate .70 results. PLS scores are based on student performance at the winter administration.

Table 19

Differential Accuracy for Screening Tasks by Grade: Hispanic-White (HW)

Grade	Parameter	df	Estimate	SE	χ^2	*p*-value
3	Intercept	1	-0.33	0.28	1.39	0.239
	PLS	1	3.69	0.65	32.12	<.001
	HW	1	-0.55	0.31	3.07	0.080
	PLS*HW	1	-1.32	0.70	3.60	0.058
4	Intercept	1	-0.66	0.33	3.98	0.046
	PLS	1	3.05	0.48	41.02	<.001
	HW	1	0.29	0.37	0.60	0.439
	PLS*HW	1	-0.56	0.55	1.04	0.307
5	Intercept	1	-0.31	0.27	1.37	0.243
	PLS	1	3.06	0.48	40.88	<.001
	HW	1	-0.39	0.30	1.63	0.202
	PLS*HW	1	-0.48	0.54	0.80	0.371
6	Intercept	1	-0.41	0.17	6.01	0.014
	PLS	1	2.62	0.34	59.29	<.001
	HW	1	-0.47	0.21	5.15	0.023
	PLS*HW	1	0.66	0.51	1.65	0.199
7	Intercept	1	-0.31	0.18	2.98	0.085
	PLS	1	3.10	0.44	48.82	<.001
	HW	1	-0.19	0.23	0.68	0.408
	PLS*HW	1	-0.37	0.55	0.44	0.509
8	Intercept	1	-0.10	0.17	0.34	0.563

	PLS	1	1.97	0.29	46.72	<.001
	HW	1	-0.72	0.22	10.20	0.001
	PLS*HW	1	0.98	0.40	5.96	0.015
9	Intercept	1	0.28	0.22	1.62	0.203
	PLS	1	2.31	0.42	30.23	<.001
	HW	1	-0.01	0.29	0.00	0.974
	PLS*HW	1	-0.59	0.52	1.28	0.258
10	Intercept	1	0.55	0.23	5.48	0.019
	PLS	1	0.99	0.30	11.05	0.001
	HW	1	-0.67	0.29	5.18	0.023
	PLS*HW	1	0.95	0.41	5.41	0.020

Note. PLS cut-off is .70. Estimates based on .85 cut-off approximate .70 results. PLS scores are based on student performance at the winter administration.

When testing for differential accuracy between ELL and non-ELL students (Table 20), a significant effect for the interaction between the PLS cut-point and ELL status existed in grade 5 (p = .01). This finding indicated that for the sample tested at the winter, non-ELL students with a PLS above the cut-point had a 90% chance of being at or above the 40th percentile on the SAT-10 compared to ELL students above the cut-point on the PLS who had a 61% chance of being at or above the 40th percentile on the SAT-10. This translates into a 29% advantage in success for non-ELL students in grade 5, but we should note that replication will be needed across multiple administrations with a larger sample to evaluate the extent to which this phenomenon continues to exist.

Table 20

Differential Accuracy for FAIR-FS Screening Tasks by Grade: English Language Learners (ELL)

Grade	Parameter	df	Estimate	SE	χ^2	*p*-value
3	Intercept	1	-0.42	0.12	12.44	<.001
	PLS	1	2.36	0.20	133.00	<.001
	ELL	1	-1.27	0.30	17.82	<.001

	PLS*ELL	1	0.71	0.66	1.15	0.284
4	Intercept	1	-0.10	0.14	0.57	0.450
	PLS	1	2.09	0.21	99.96	<.001
	ELL	1	-1.00	0.30	11.23	<.001
	PLS*ELL	1	0.24	0.89	0.07	0.788
5	Intercept	1	-0.50	0.13	14.52	<.001
	PLS	1	2.72	0.21	168.37	<.001
	ELL	1	-0.38	0.24	2.46	0.117
	PLS*ELL	1	-1.41	0.54	6.68	0.010
6	Intercept	1	-0.47	0.10	22.46	<.001
	PLS	1	2.46	0.21	134.43	<.001
	ELL	1	-1.37	0.25	29.01	<.001
	PLS*ELL	1	-0.63	0.79	0.63	0.426
7	Intercept	1	-0.08	0.11	0.59	0.441
	PLS	1	2.47	0.24	108.98	<.001
	ELL	1	-1.56	0.27	34.34	<.001
	PLS*ELL	1	-0.01	0.74	0.00	0.991
8	Intercept	1	-0.14	0.10	1.70	0.192
	PLS	1	2.11	0.18	134.92	<.001
	ELL	1	-1.74	0.28	40.22	<.001
	PLS*ELL	1	1.37	0.76	3.28	0.070
9	Intercept	1	0.29	0.13	5.04	0.025
	PLS	1	1.93	0.22	80.32	<.001
	ELL	1	-0.59	0.34	3.00	0.083

	PLS*ELL	1	-1.23	0.91	1.81	0.178
10	Intercept	1	0.20	0.13	2.49	0.114
	PLS	1	1.54	0.18	75.19	<.001
	ELL	1	-1.16	0.35	11.19	0.001
	PLS*ELL	1	-0.63	0.59	1.12	0.291

Note. PLS cut-off is .70. Estimates based on .85 cut-off approximate .70 results. PLS scores are based on student performance at the winter administration.

When testing for differential accuracy between FRL and non- FRL students (Table 21), a significant effect for the interaction between the PLS cut-point and FRL status existed in grade 10 (p = .002). This finding indicated that for the sample tested at the winter, non- FRL students with a PLS above the cut-point had a 91% chance of being at or above the 40th percentile on the SAT-10 compared to FRL students above the cut-point on the PLS who had a 75% chance of being at or above the 40th percentile on the SAT-10. This translates into a 16% advantage in success for non-FRL students in grade 10, but we should note that replication will be needed across multiple administrations with a larger sample to evaluate the extent to which this phenomenon continues to exist.

Table 21

Differential Accuracy for Screening Tasks by Grade: Free or Reduced Price Lunch (FRL)

Grade	Parameter	df	Estimate	SE	χ^2	*p*-value
3	Intercept	1	0.59	0.32	3.56	0.059
	PLS	1	3.11	0.75	17.16	<.001
	FRL	1	-1.45	0.34	18.57	<.001
	PLS*FRL	1	-0.65	0.78	0.70	0.403
4	Intercept	1	1.00	0.41	5.83	0.016
	PLS	1	1.58	0.54	8.63	0.003
	FRL	1	-1.50	0.43	11.99	0.001
	PLS*FRL	1	0.66	0.58	1.29	0.257
5	Intercept	1	-0.17	0.34	0.24	0.623

	PLS	1	2.77	0.47	34.72	<.001
	FRL	1	-0.50	0.36	1.99	0.159
	PLS*FRL	1	-0.22	0.51	0.19	0.664
6	Intercept	1	-0.54	0.19	7.67	0.006
	PLS	1	2.95	0.38	61.37	<.001
	FRL	1	-0.27	0.22	1.53	0.216
	PLS*FRL	1	-0.57	0.45	1.64	0.200
7	Intercept	1	0.29	0.21	1.79	0.180
	PLS	1	2.63	0.44	36.49	<.001
	FRL	1	-0.90	0.24	13.97	0.000
	PLS*FRL	1	-0.10	0.51	0.04	0.836
8	Intercept	1	-0.01	0.19	0.00	0.948
	PLS	1	2.22	0.30	55.92	<.001
	FRL	1	-0.64	0.22	8.52	0.004
	PLS*FRL	1	0.19	0.37	0.26	0.611
9	Intercept	1	0.45	0.21	4.71	0.030
	PLS	1	1.99	0.33	36.53	<.001
	FRL	1	-0.37	0.25	2.13	0.144
	PLS*FRL	1	-0.13	0.42	0.10	0.752
10	Intercept	1	0.08	0.18	0.22	0.642
	PLS	1	2.21	0.27	65.32	<.001
	FRL	1	-0.10	0.23	0.18	0.675
	PLS*FRL	1	-1.08	0.35	9.64	0.002

Note. PLS cut-off is .70. Estimates based on .85 cut-off approximate .70 results. PLS scores are based on student performance at the winter administration.

Construct Validity

Construct validity describes how well scores from an assessment measure the construct it is intended to measure. Components of construct validity include convergent validity, which can be evaluated by testing relations between a developed assessment and another related assessment, and discriminant validity, which can be evaluated by correlating scores from a developed assessment with an unrelated assessment. The goal of the former is to yield a high association which indicates that the developed measure converges, or is empirically linked to, the intended construct. The goal of the latter is to yield a lower association, which indicates that the developed measure is unrelated to a particular construct of interest.

Convergent validity*.* Data was collected in two large school districts in central Florida with four elementary schools, three middle schools, and two high schools. A total of 1,825 students in grades 3 through 10 were administered the four tasks in the FAIR-FS and gold standard clinical norm-referenced assessments of word reading (Test of Word Reading Efficiency – 2, Wagner, Torgesen, & Rashotte, 2012), vocabulary (Peabody Picture Vocabulary Test – 4, Dunn & Dunn, 2007), and syntax (the Grammaticality Judgment Test of the Comprehensive Assessment of Spoken Language, Carrow-Woolfolk, 2008).

Students' abilities to derive word meanings receptively was measured by the VKT and the Peabody Picture Vocabulary Test-4 (PPVT-4; Dunn & Dunn, 2007). The PPVT-4 is used frequently as a normative measure and as a diagnostic. The PPVT-4 requires students to point to a picture, from a group of four pictures, which best represents a word spoken by the examiner. The PPVT-4 manual reports high reliability, with internal consistency reliability ranging from .92 to .98. The PPVT-4 also demonstrates high convergent validity to other measures, with correlations ranging from .80 to .83 with the Expressive Vocabulary Test (Williams, 2007) and correlations with the Clinical Evaluation of Language Fundamentals (Semel, Wiig, & Secord, 2003) ranging from .67 to .79.

Students' abilities to use the structure of sentences to comprehend the sentences' meaning was measured by the SKT and the Grammaticality Judgment subtest (GJT) of the Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk, 2008). The CASL is most frequently used by speech language pathologists to determine instructional/therapy goals for students with diagnostic weaknesses in language skills such as syntax. In the GJT, students were orally presented sentences with and without grammatical errors and asked indicate whether or not there were errors. The items have an additional component asking students to fix any perceived errors in the sentence without changing its meaning. The GJT subtest has high internal consistency reliability ranging from .85 to .94 and high criterion-related validity with other oral language assessments within the CASL. The manual reports that, after correcting for variability between norm groups, the GJT correlates to the Listening Comprehension and Oral Expression Scales (Carrow-Woolfolk, 1995) Oral Composite score at .75.

Word recognition was measured by the WRT and compared to performance of a measure of decoding fluency, the Sight Word Efficiency and Phonemic Decoding Efficiency subtests of the Test of Word

Reading Efficiency-2 (TOWRE-2; Wagner, Torgesen, & Rashotte, 2012). The TOWRE-2 was designed to monitor the progress of students receiving additional instruction for weaknesses in word reading abilities and has demonstrated discrimination between low-performing students with language and reading disabilities (Wagner, Togesen, & Rashotte, 2012). When administering this assessment, the examiner asks students to read nonwords and sight words aloud as quickly as possible within 45 seconds. The alternate-forms reliability coefficient ranges from .82-.94 and average test–retest coefficients amongst forms exceeds .90. Correlations with other measures of word reading is high, such as the Letter-Word Identification subtest of the Woodcock-Johnson III (r = .76; Woodcock, McGrew, & Mather, 2001), reading fluency (r = .91) on the Gray Oral Reading Test-4th ed. (GORT-4; Wiederholt & Bryant, 2001), Test of Silent Contextual Reading Fluency (TOSCRF; Hammill, Wiederholt, & Allen, 2006; r = .75), and the Woodcock Reading Mastery Test–Revised (WRMT-R; Woodcock, 1987) Passage Comprehension (r = .88).

Relations between the FAIR-FS Reading Comprehension Task and the SAT-10 Reading Comprehension are found in Table 16. Correlations in Table 22 demonstrate moderate associations exist between the FAIR-FS Vocabulary Knowledge Task and the PPVT-IV. The average correlation across grade levels is .52 with a range of .47 to .67. Correlations between the FAIR-FS Word Recognition Task and the TOWRE Real Word component of the TOWRE demonstrated moderate associations as well. The average correlation across grade levels is .33 with a range of .24 to .49. Correlations between the FAIR-FS Word Recognition Task and the TOWRE Non-Word component of the TOWRE were moderate. The average correlation across grade levels was .38 with a range of .30 to .47. Correlations between the FAIR-FS Syntax Knowledge Task and the GJT were moderate. The average correlation across grade levels was .49, with a range of .37 to .61.

Table 22

Correlations between FAIR-FS scores and the PPVT-IV, GJT, and TOWRE

Grade	N	FAIR-FS Task	PPVT-IV	GJT	TOWRE Real Word	TOWRE Non-Word
3	251	Vocabulary Knowledge	**0.47**	0.40	0.37	0.29
		Syntax Knowledge	0.54	**0.49**	0.34	0.28
		Word Recognition	0.27	0.31	**0.42**	**0.43**
4	161	Vocabulary Knowledge	**0.56**	0.57	0.50	0.44
		Syntax Knowledge	0.60	**0.61**	0.35	0.33
		Word Recognition	0.36	0.40	**0.45**	**0.45**
5	167	Vocabulary Knowledge	**0.61**	0.51	0.35	0.39
		Syntax Knowledge	0.56	**0.47**	0.33	0.32
		Word Recognition	0.22	0.10	**0.24**	**0.30**
6	113	Vocabulary Knowledge	**0.62**	0.53	0.41	0.44
		Syntax Knowledge	0.52	**0.44**	0.20	0.20
		Word Recognition	0.36	0.26	**0.49**	**0.47**
7	72	Vocabulary Knowledge	**0.58**	0.50	0.43	0.33
		Syntax Knowledge	0.50	**0.49**	0.30	0.28
		Word Recognition	0.34	0.31	**0.46**	**0.51**
8	71	Vocabulary Knowledge	**0.50**	0.53	0.36	0.45
		Syntax Knowledge	0.74	**0.51**	0.33	0.47
		Word Recognition	0.41	0.45	**0.28**	**0.46**
9	227	Vocabulary Knowledge	**0.65**	0.55	0.27	0.29
		Syntax Knowledge	0.35	**0.37**	0.25	0.27
		Word Recognition	0.39	0.25	**0.35**	**0.43**
10	169	Vocabulary Knowledge	**0.67**	0.61	0.36	0.44
		Syntax Knowledge	0.52	**0.56**	0.34	0.38
		Word Recognition	0.40	0.40	**0.28**	**0.36**

Note. PPVT-IV = Peabody Picture Vocabulary Task – 4th Edition; GJT = Grammaticality Judgment Task, TOWRE = Test of Word Reading Efficiency.

A secondary analysis of convergent validity evaluated the extent to which the correlations between the FAIR-FS and the PPVT-IV, GJT, and TOWRE tasks varied dependent on one's level of ability. Because traditional correlations are representative of average associations, it is possible that the average does not best characterize relations for students with low, average, and high ability levels. For example, it is plausible that at low levels of the GJT, a stronger correlation exists between the GJT and the FAIR-FS Syntax Knowledge compared to a weaker correlation at higher levels of the GJT. Because the GJT is a clinical measure of syntax knowledge, it is designed for students who are supposed to be deficient in this skill. The GJT is not typically administered to students with average or high syntax skills; therefore, reporting the average correlation between scores on the GJT and the FAIR-FS Syntactic Knowledge could mask a stronger association for students with poor syntax skills. Typical regression models are ill-

equipped to test for differential correlations across the range of scores for an outcome variable. Rather, quantile regression (Koenker & Bassett, 1978; Petscher & Logan, 2014; Petscher, Logan, & Zhou, 2013) is suitable to estimating the correlation between measures conditional on performance of the outcome. In this manner we tested the extent to which: 1) the correlation between the FAIR-FS Vocabulary Knowledge and PPVT-IV varied for students with low, average, and high PPVT-IV scores; 2) the correlation between the FAIR-FS Word Recognition and TOWRE-Real Word varied for students with low, average, and high TOWRE Real Word scores; 3) the correlation between the FAIR-FS Word Recognition and TOWRE Non-Word varied for students with low, average, and high TOWRE Non-Word scores; and 4) the correlation between the FAIR-FS Syntactic Knowledge and GJT varied for students with low, average, and high GJT scores.

Figures from the quantile correlation analyses are reported in Appendices D-G. The quantile correlations between FAIR-FS Vocabulary Knowledge and the PPVT-IV (Appendix D) show that in general the correlations between the two assessments are more strongly related for students who performed lower in PPVT-IV. The implication is that lower performance on the PPVT-IV is correlated with low performance on the Vocabulary Knowledge task. At higher levels of the PPVT-IV the correlation is still moderate but less than that observed at the lower level of PPVT-IV. To better capture the nature of the relations between the variables, Table 23 provides a summary of the average correlation between the two tasks by grade for students who are low on the PPVT-IV (i.e., <40th quantile/percentile), average (40th -60th quantile/percentile) and high (> 60th quantile/percentile). The quantile correlations demonstrate a trend that higher correlations between the measures are observed for students who score low or average on the PPVT-IV. A similar trend is generally observed for the FAIR-FS Word Recognition Task in its relation to the two TOWRE outcomes (Appendix E and F; Table 23) as well as for the Syntactic Knowledge Task (Appendix G; Table 23).

Table 23

Average correlations within ranges of quantiles/percentiles by grade and task

FAIR-FS Task	Outcome	Grade	Quantile/Percentile Range <40	40-60	>60
Vocabulary Knowledge	PPVT-IV	3	0.60	0.48	0.40
		4	0.60	0.50	0.42
		5	0.66	0.67	0.52
		6	0.67	0.58	0.54
		7	0.66	0.63	0.55
		8	0.52	0.34	0.25
		9	0.72	0.56	0.51
		10	0.72	0.70	0.54
Word Recognition	TOWRE Real Word	3	0.47	0.47	0.34
		4	0.44	0.41	0.40
		5	0.19	0.19	0.16
		6	0.54	0.48	0.49
		7	0.53	0.45	0.40
		8	0.11	0.28	0.38
		9	0.31	0.29	0.43
		10	0.19	0.31	0.37
Word Recognition	TOWRE Non-Word	3	0.45	0.48	0.35
		4	0.50	0.47	0.35
		5	0.39	0.31	0.27
		6	0.57	0.38	0.36
		7	0.67	0.38	0.33
		8	0.55	0.37	0.38
		9	0.48	0.41	0.29
		10	0.52	0.33	0.18
Syntactic Knowledge	GJT	3	0.44	0.52	0.52
		4	0.66	0.58	0.58
		5	0.50	0.52	0.40
		6	0.50	0.48	0.37
		7	0.71	0.41	0.50
		8	0.70	0.48	0.30
		9	0.39	0.38	0.47
		10	0.61	0.55	0.52

Note. PPVT-IV = Peabody Picture Vocabulary Task – 4th Edition; GJT = Grammaticality Judgment Task, TOWRE = Test of Word Reading Efficiency.

Discriminant validity. Discriminant validity was evaluated by estimating correlations between the FAIR-FS tasks and variables that should not be related to measures of reading: sex and birthdate (Table 24). Results indicated that weak associations were generally observed across grade levels.

Table 24

Correlations between FAIR-FS tasks and birthdate/sex

Grade	Task	Birthdate	Sex
3	Vocabulary Knowledge	0.10	0.11
	Word Recognition	0.09	0.08
	Reading Comprehension	0.11	0.22
	Syntax Knowledge	0.04	0.08
4	Vocabulary Knowledge	0.16	-0.02
	Word Recognition	0.21	0.03
	Reading Comprehension	0.14	0.14
	Syntax Knowledge	0.09	0.04
5	Vocabulary Knowledge	0.13	-0.12
	Word Recognition	0.02	-0.01
	Reading Comprehension	0.23	0.13
	Syntax Knowledge	0.17	-0.12
6	Vocabulary Knowledge	0.26	-0.20
	Word Recognition	0.14	-0.01
	Reading Comprehension	0.28	-0.20
	Syntax Knowledge	0.23	-0.20
7	Vocabulary Knowledge	0.01	-0.12
	Word Recognition	0.20	0.00
	Reading Comprehension	0.12	-0.06
	Syntax Knowledge	0.22	0.05
8	Vocabulary Knowledge	0.01	-0.26
	Word Recognition	0.12	-0.13
	Reading Comprehension	0.09	0.04
	Syntax Knowledge	0.12	-0.16
9	Vocabulary Knowledge	0.15	-0.10
	Word Recognition	0.12	-0.10
	Reading Comprehension	0.12	0.01
	Syntax Knowledge	0.18	0.12
10	Vocabulary Knowledge	0.20	0.04
	Word Recognition	0.14	0.02
	Reading Comprehension	0.18	0.10
	Syntax Knowledge	0.20	0.17

References

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology*, *103*(2), 429-441.

Carlisle, J. F., Stone, C. A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly, 40*, 428-449.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language. Circle Pines*, MN: AGS.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second ed.). Hillsdale: Lawrence Erlbaum Associates.

Crosson, A. C., Lesaux, N. K. (2013). Connectives: Fitting another piece of the vocabulary instruction puzzle. *The Reading Teacher, 67*(3), 193-200.

Dunn, L.M., & Dunn, D.M. (2007). *Peabody Picture Vocabulary Test-IV*. Circle Pines, MN: American Guidance Service.

Ehri, L. C., Nunes, S., Stahl, S., & Willows, D. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*, 393-447.

Ehri, L. C., Nunes, S., Willows, D., Schuster, B., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the national Reading Panel's meta-analysis. *Reading Research Quarterly, 36*, 250-287.

Florida Department of Education (2009-2014). Florida Assessments for Instruction in Reading (FAIR). Tallahassee, FL: Author.

Foorman, B. R. (2009). Text difficulty in reading assessment. In E.H. Hiebert (Ed.), *Reading more, reading better* (pp. 231-247.) New York, NY: Guilford.

Foorman, B. R., & Connor, C. (2011). Primary reading. In M. Kamil, P. D. Pearson, & E. Moje (Eds.), *Handbook on reading research* (Vol. 4, pp. 136–156). New York, NY: Taylor and Francis.

Foorman, B., & Herrera, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). The Structure of Oral Language and Reading and Their Relation to Comprehension in grades K-2. *Reading and Writing*. Advance online publication. doi: 10.1007/s11145-015-9544-5.

Foorman, B., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining General and Specific Factors in the Dimensionality of Oral Language and Reading in 4th-10th Grades. *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/edu0000026.

Foorman, B. R., & Petscher, Y. (2010a). *Summary of the predictive relationship between the FAIR and the FCAT in grades 3-10*. Tallahassee, FL: Florida Center for Reading Research.

Foorman, B. R., & Petscher, Y. (2010b). *The unique role of the FAIR Broad Screen in predicting FCAT Reading Comprehension*. Tallahassee, FL: Florida Center for Reading Research.

Foorman, B. R., Petscher, Y., & Bishop, M. D. (2012). The incremental variance of morphological knowledge to reading comprehension in grades 3-10 beyond prior reading comprehension, spelling, and text reading efficiency. *Learning and Individual Differences, 22*, 792-798.

Hammill, D. D., Wiederholt, J. L., & Allen, E. A. (2006). *TOSCRF: Test of silent contextual reading fluency: Examiner's manual.* Austin, TX: Pro-Ed.

Harvill, L.M. (2005). Standard error of measurement. *Educational Measurement: Issues and Practice, 10,* 33-41.

Hattie, J., Krakowski, K., Rogers, H.J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*, 1-14

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46,* 33-50.

Meade, A. W. (2010). A taxonomy of effect sizes for the differential functioning of items and scales. *Journal of Applied Psychology, 95*, 728-743.

McDonald, R.P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research, 14*, 21-38.

Muthén, B., & Muthén, L. (2008). *Mplus User's Guide*. Los Angeles, CA: Muthén and Muthén.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel.* Washington, DC: National Institute for Literacy. Retrieved from: http://lincs.ed.gov/publications/pdf/NELPReport09.pdf

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Department of Health and Human Services.

National Research Council (1998). *Preventing reading difficulties in young children*. Committee on the Prevention of Reading Difficulties in Young Children, Committee on Behavioral and Social Science and Education, C.E. Snow, M.S. Burns, & P. Griffin, eds. Washington, D.C.: National Academy Press.

Nelson, J., Perfetti, C., Liben, D., Liben, M. (2012). Measures of text difficulty: Testing their predictive value for grade levels and student performance. Retrieved from: http://www.ccsso.org/Documents/2012/Measures%20ofText%20Difficulty_final.2012.pdf

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd Ed.). New York: McGraw-Hill.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22-37. DOI: 10.1080/10888438.2013.827687

Petscher, Y., & Foorman, B. R. (2011). *Summary of the predictive relationship between the FAIR and the FCAT in grades 3-10*. Tallahassee, Fl: Florida Center for Reading Research.

Petscher, Y., Kim, Y.S., & Foorman, B.R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the RTI framework. *Assessment for Effective Intervention, 36,* 158-166.

Petscher, Y., Logan, J.A.R., & Zhou, C. (2013). Extending conditional means modeling: An introduction to quantile regression (pp. 3-33). In Y. Petscher, C. Schatschneider, & D.L. Compton (Eds.) *Applied quantitative analysis in education and social sciences*. New York: Routledge.

Petscher, Y., & Logan, J.A.R. (2014). Quantile regression in the study of Developmental Sciences. *Child Development, 85,* 861-881. Doi: 10.1111/cdev.12190.

RAND Reading Study Group (2002). *Reading for understanding*. Santa Monica, CA: RAND Corporation.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31-74.

Schatschneider, C., Petscher, Y., & Williams, K. M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know (pg. 304-317). In L. Justice & C. Vukelic (Eds.). *Every moment counts: Achieving excellence in preschool language and literacy instruction*. New York: Guilford Press.

Semel, E., Wiig, E. & Secord, W. (2003). *Clinical evaluation of language fundamentals -4 (CELF-4).* San Antonio, TX: PyschCorp.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Stout, W.F. (1990). A new item response theory modeling approach and applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81*, 209-219.

Torgesen, J.K., Wagner, R.K., & Rashotte, C.A. (2012). *Test of Word Reading Efficiency, 2nd Edition*. Austin, TX: Pro-ED, Inc. Wiederholt J. L & Bryant B. R. (2001) *Gray Oral Reading Test-IV.* Austin; TX: Pro-ED, Inc.

Williams, K. T. (2007). *Expressive Vocabulary Test, Second Edition. (2nd ed.)* Minneapolis, MN: Pearson Assessments.

Woodcock, R. W. (1987). *Woodcock reading mastery tests–Revised*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Test.* Riverside PublishingCompany. Itasca, IL

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide.* New York: Touchstone Applied Science Associates. Inc., Brewster.

Ziegler, J. C., Stone, G. O., & Jacobs, A. M. (1997). What's the pronunciation for _OUGH and the spelling for /u/? A database for computing feedforward and feedback inconsistency in English. *Behavior Research Methods, Instruments, & Computers, 29*, 600-618.

Appendix A: G3-G12 Weights

Table A1

Population values for each grade for each of the sixteen demographic groups

			Grade							
Race	FRL	ELL	3	4	5	6	7	8	9	10
White	Yes	Yes	0.00	0.20	0.64	0.25	0.33	0.36	0.17	0.43
White	Yes	No	18.11	17.52	17.26	17.69	16.80	16.37	15.24	13.48
White	No	Yes	0.09	0.30	0.09	0.13	0.00	0.07	0.25	0.14
White	No	No	22.02	23.22	23.61	23.69	25.00	25.95	27.56	29.39
Black	Yes	Yes	0.18	0.30	0.27	0.19	0.47	0.43	0.25	0.57
Black	Yes	No	19.75	18.72	18.80	18.88	18.20	17.53	16.49	15.55
Black	No	Yes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Black	No	No	3.00	3.20	3.36	3.75	4.20	4.50	5.83	6.56
Hispanic	Yes	Yes	6.82	5.51	7.27	7.75	7.07	6.79	2.83	3.85
Hispanic	Yes	No	16.65	17.42	15.26	14.38	14.53	14.37	16.49	14.41
Hispanic	No	Yes	0.18	0.20	0.18	1.00	0.40	0.79	0.67	0.86
Hispanic	No	No	6.73	6.81	6.81	6.06	6.93	7.01	8.33	8.92
Other	Yes	Yes	0.00	0.30	0.09	0.25	0.53	0.21	0.25	0.36
Other	Yes	No	3.46	3.21	3.36	3.00	2.60	2.57	2.41	2.07
Other	No	Yes	0.09	0.10	0.00	0.06	0.07	0.07	0.08	0.00
Other	No	No	2.91	3.00	3.00	2.94	2.87	2.93	3.16	3.42

Note. Not all race/ethnicity subgroups are represented due to limited information provided when evaluating interactions among (i.e., White, Black, Hispanic, Other), free/reduced lunch status (eligible or ineligible), and English language learner (identified or not identified). Students in grades 11 and 12 use the grade 10 distribution of ability scores. FRL = Free/reduced price lunch. ELL = English language learners.

Table A2

Sample weight values for Reading Comprehension Task

			Grade							
Race	FRL	ELL	3	4	5	6	7	8	9	10
White	Yes	Yes	0.00	0.77	1.16	1.09	1.22	2.00	1.13	1.65
White	Yes	No	0.91	1.04	1.04	1.32	1.26	1.34	1.10	1.08
White	No	Yes	0.41	1.58	1.29	0.57	0.00	0.44	1.92	1.08
White	No	No	0.58	0.53	0.52	0.67	0.71	0.72	0.97	0.96
Black	Yes	Yes	1.64	2.73	2.45	0.61	0.87	0.61	0.45	1.04
Black	Yes	No	2.08	2.11	2.06	1.31	1.18	1.16	0.85	0.92
Black	No	Yes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Black	No	No	0.86	0.92	1.09	1.04	1.40	1.16	1.34	1.17
Hispanic	Yes	Yes	1.93	1.92	1.96	1.41	1.39	1.42	1.04	1.23
Hispanic	Yes	No	1.83	1.93	2.03	0.91	0.96	0.98	0.94	0.92
Hispanic	No	Yes	0.33	0.54	0.62	1.23	0.59	1.08	1.56	1.37
Hispanic	No	No	1.05	1.28	1.39	1.24	1.31	1.17	1.52	1.41
Other	Yes	Yes	0.00	1.00	0.23	0.96	1.39	0.72	1.39	1.24
Other	Yes	No	1.11	1.10	0.99	1.15	1.03	1.11	0.78	0.69
Other	No	Yes	0.24	0.29	0.00	0.46	0.64	0.44	0.80	0.00
Other	No	No	0.53	0.60	0.71	1.31	1.01	1.16	0.91	0.81

Note. Not all race/ethnicity subgroups are represented due to limited information provided when evaluating interactions among (i.e., White, Black, Hispanic, Other), free/reduced lunch status (eligible or ineligible), and English language learner (identified or not identified). Students in grades 11 and 12 use the grade 10 distribution of ability scores. FRL = Free/reduced price lunch. ELL = English language learners.

Table A3

Sample weight values for Vocabulary Knowledge Task

			Grade							
Race	FRL	ELL	3	4	5	6	7	8	9	10
White	Yes	Yes	0.00	2.00	0.64	0.25	0.33	0.36	0.17	0.43
White	Yes	No	0.69	0.67	0.66	0.89	0.87	0.72	0.69	0.77
White	No	Yes	9.00	0.30	0.09	0.13	0.00	0.07	0.25	0.14
White	No	No	0.84	0.81	0.82	1.06	1.01	0.90	1.01	0.89
Black	Yes	Yes	0.90	1.67	1.50	0.19	2.76	2.53	0.68	0.57
Black	Yes	No	1.77	2.01	1.52	1.14	1.00	1.18	1.07	1.06
Black	No	Yes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Black	No	No	1.00	0.96	0.97	0.65	0.72	0.93	1.44	1.08
Hispanic	Yes	Yes	2.85	10.40	5.39	7.83	6.04	5.18	2.16	3.16
Hispanic	Yes	No	0.93	0.88	0.89	0.56	0.71	0.71	0.87	1.02
Hispanic	No	Yes	18.00	1.11	0.95	5.88	2.35	4.65	5.58	0.86
Hispanic	No	No	1.35	1.55	1.36	1.41	1.66	1.74	1.88	1.31
Other	Yes	Yes	0.00	3.00	0.47	0.25	0.53	0.21	2.08	0.36
Other	Yes	No	0.96	1.14	1.34	1.52	1.04	1.53	0.89	0.92
Other	No	Yes	9.00	0.56	0.00	0.06	0.07	0.07	0.08	0.00
Other	No	No	0.86	0.71	1.20	1.37	0.95	2.90	0.95	0.85

Note. Not all race/ethnicity subgroups are represented due to limited information provided when evaluating interactions among (i.e., White, Black, Hispanic, Other), free/reduced lunch status (eligible or ineligible), and English language learner (identified or not identified). Students in grades 11 and 12 use the grade 10 distribution of ability scores. FRL = Free/reduced price lunch. ELL = English language learners.

Table A4

Sample weight values for Word Recognition Task

			Grade							
Race	FRL	ELL	3	4	5	6	7	8	9	10
White	Yes	Yes	0.00	1.18	0.64	0.25	0.33	0.36	1.89	0.43
White	Yes	No	1.71	1.63	1.60	2.45	2.23	2.45	2.82	3.56
White	No	Yes	0.09	0.30	0.09	0.13	0.00	0.44	1.32	0.14
White	No	No	0.52	0.51	0.54	0.55	0.49	0.50	0.59	0.48
Black	Yes	Yes	0.18	0.30	0.27	0.19	2.94	0.43	2.78	6.33
Black	Yes	No	0.83	0.84	0.87	0.67	0.84	1.01	0.72	1.00
Black	No	Yes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Black	No	No	0.32	0.38	0.29	0.36	0.41	0.39	0.48	0.60
Hispanic	Yes	Yes	45.47	16.21	51.93	75.00	70.00	6.79	2.83	10.69
Hispanic	Yes	No	9.05	14.64	6.63	9.59	12.01	11.98	16.49	11.44
Hispanic	No	Yes	1.20	1.18	0.18	1.00	0.40	4.94	3.53	3.19
Hispanic	No	No	2.58	2.66	2.37	2.82	2.83	3.92	2.74	3.96
Other	Yes	Yes	0.00	0.88	0.64	1.67	1.61	0.64	0.89	0.36
Other	Yes	No	1.07	1.45	2.92	1.09	1.14	1.31	1.70	2.30
Other	No	Yes	0.20	0.59	0.00	0.06	0.44	0.21	0.89	0.00
Other	No	No	0.57	0.53	0.55	0.83	1.17	0.49	0.51	0.92

Note. Not all race/ethnicity subgroups are represented due to limited information provided when evaluating interactions among (i.e., White, Black, Hispanic, Other), free/reduced lunch status (eligible or ineligible), and English language learner (identified or not identified). Students in grades 11 and 12 use the grade 10 distribution of ability scores. FRL = Free/reduced price lunch. ELL = English language learners.

Table A5

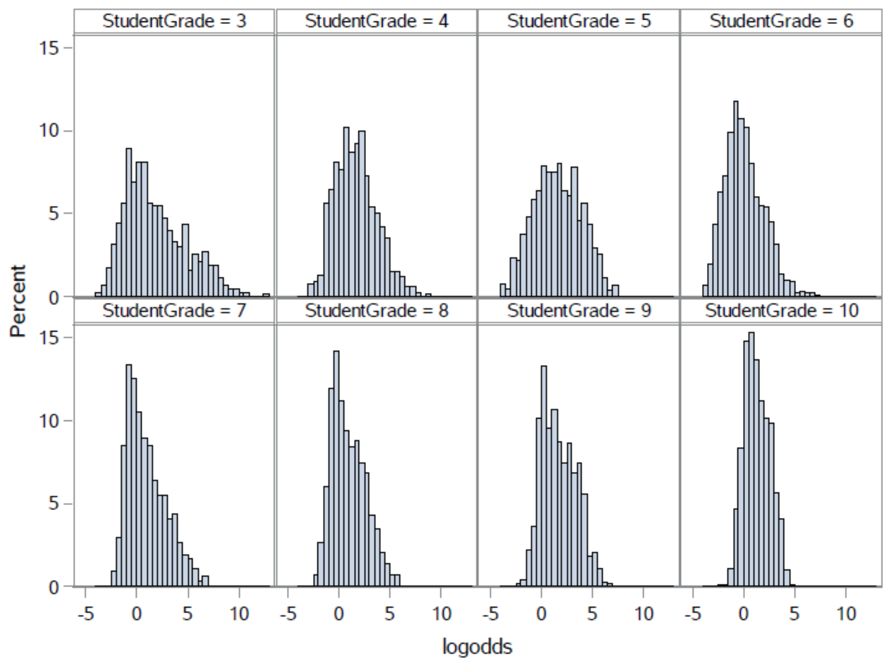
Sample weight values for Syntactic Knowledge Task

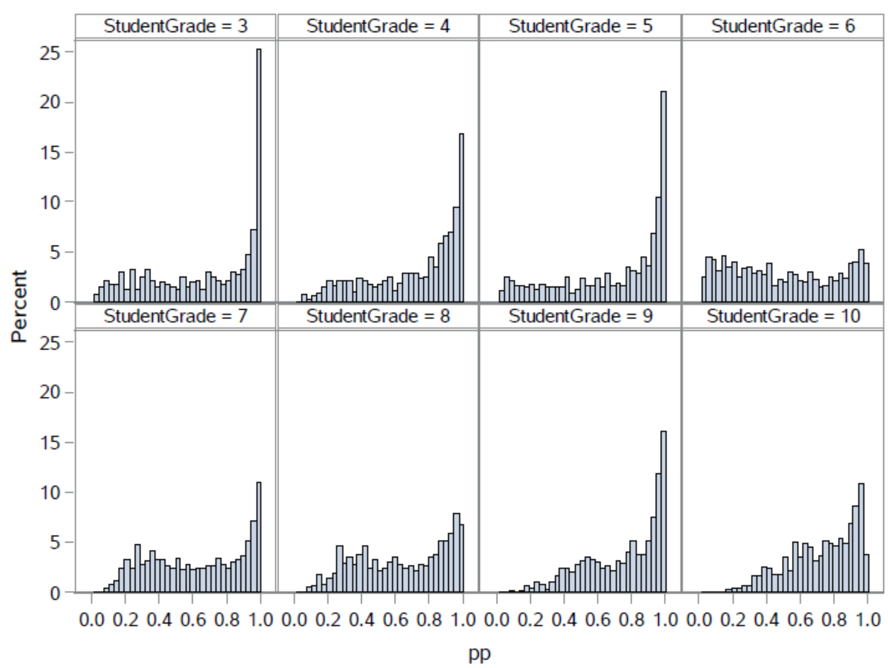
			Grade							
Race	FRL	ELL	3	4	5	6	7	8	9	10
White	Yes	Yes	0.00	1.67	1.00	1.00	1.00	36.00	17.00	43.00
White	Yes	No	2.39	2.14	2.27	2.31	3.23	14.36	14.65	12.96
White	No	Yes	0.29	1.00	1.00	1.00	0.00	7.00	25.00	14.00
White	No	No	0.50	0.47	0.43	0.39	0.37	0.33	0.37	0.38
Black	Yes	Yes	0.10	0.14	0.33	0.23	2.94	43.00	2.78	57.00
Black	Yes	No	0.83	0.98	1.15	1.08	1.26	2.12	1.31	1.35
Black	No	Yes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Black	No	No	0.46	0.55	0.51	0.59	0.58	0.73	0.89	0.93
Hispanic	Yes	Yes	9.34	2.36	5.08	13.36	70.70	679.00	283.00	385.00
Hispanic	Yes	No	3.27	3.83	3.32	4.61	29.65	29.33	24.98	120.08
Hispanic	No	Yes	0.43	1.67	1.80	100.00	4.00	79.00	67.00	86.00
Hispanic	No	No	2.31	3.89	2.67	3.50	4.28	14.31	14.61	38.78
Other	Yes	Yes	0.00	2.50	0.29	2.08	3.31	1.31	25.00	36.00
Other	Yes	No	1.23	1.06	2.35	1.52	1.78	1.76	4.23	9.00
Other	No	Yes	0.17	0.83	0.00	0.50	0.44	0.44	0.89	0.00
Other	No	No	1.12	0.99	0.89	1.16	1.61	1.39	0.88	2.28

Note. Not all race/ethnicity subgroups are represented due to limited information provided when evaluating interactions among (i.e., White, Black, Hispanic, Other), free/reduced lunch status (eligible or ineligible), and English language learner (identified or not identified). Students in grades 11 and 12 use the grade 10 distribution of ability scores. FRL = Free/reduced price lunch. ELL = English language learners.

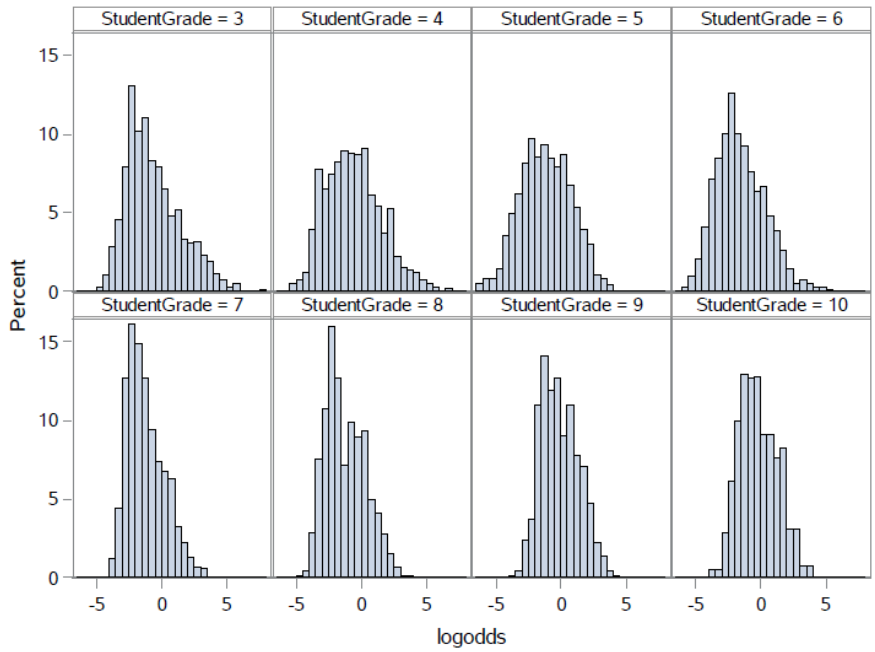
Note that Table A1 should be used with Tables A2 through A5. Large sample weights reflect subgroups which needed to be weighted more in the analyses; however, a large value does not necessarily indicate gross under-sampling. For example, Table A.5 highlights that Hispanic students who are FRL and ELL have large weights in grades 8-10 (e.g., 679, 283, and 385). Note also that Table A1 shows that Hispanic students who are FRL and ELL constitute only 6.79% of the state population in grade 8. Thus, the large sample weight reflects the need to weight the smaller sample by a factor of 679 so that it can adequately reflect the state population at an appropriate level.

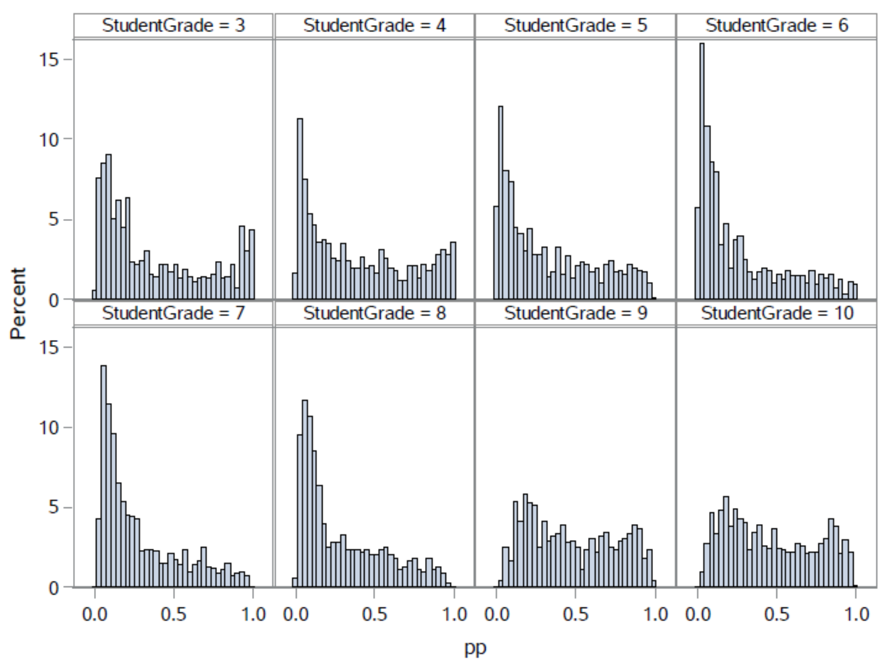
Appendix B: Distribution of the Log Odds and Predicted Probability of Success on the SAT-10 at the 40th Percentile



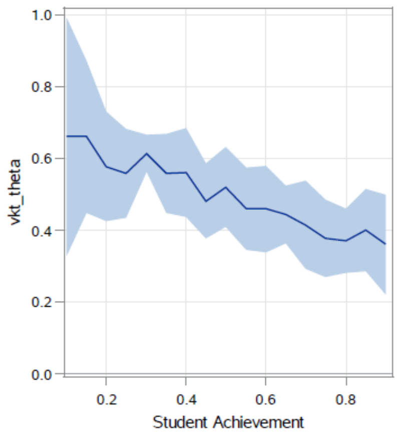


Appendix C: Distribution of the Log Odds and Predicted Probability of Success on the SAT-10 at the 70th Percentile

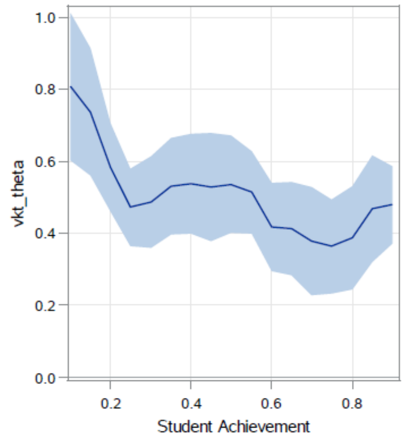




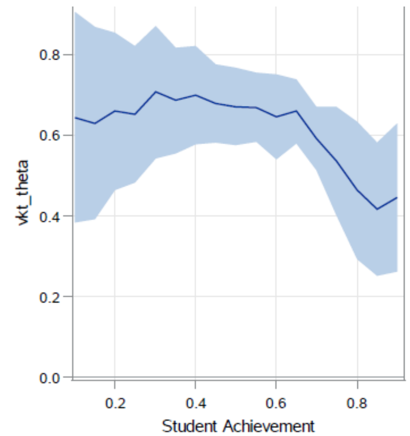
Appendix D: Quantile Correlations between FAIR-FS Vocabulary Knowledge and PPVT-IV



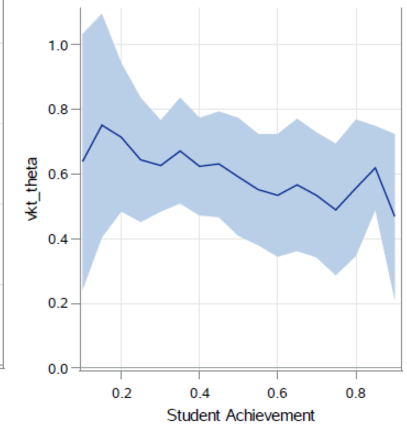
Grade 3



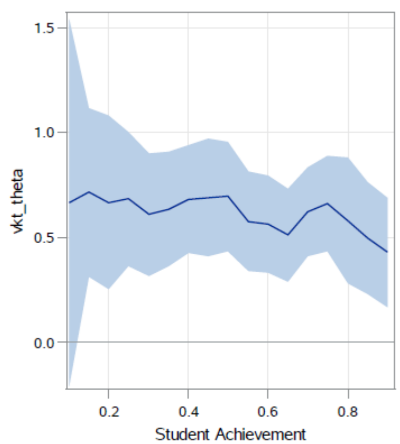
Grade 4



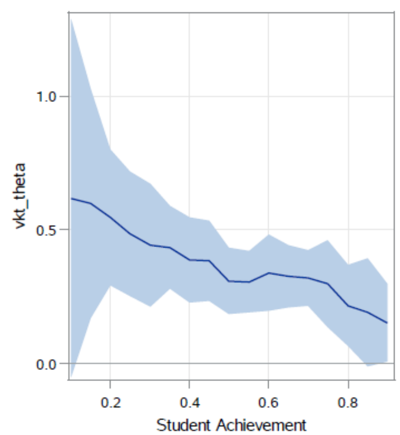
Grade 5



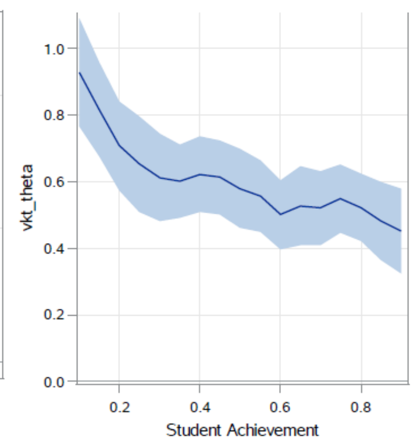
Grade 6



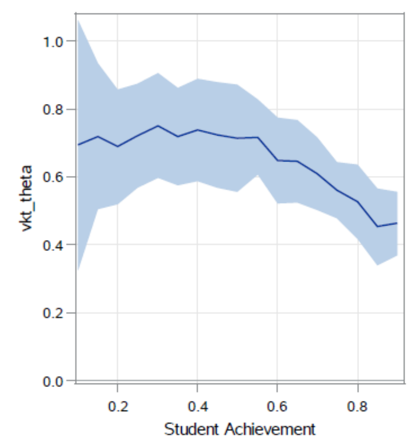
Grade 7



Grade 8

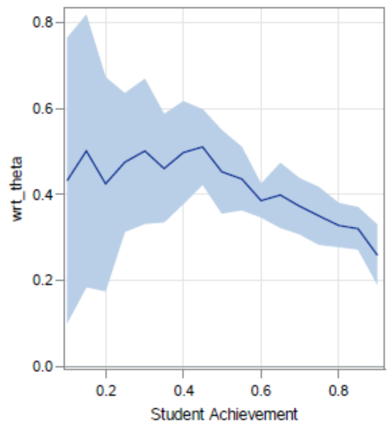


Grade 9

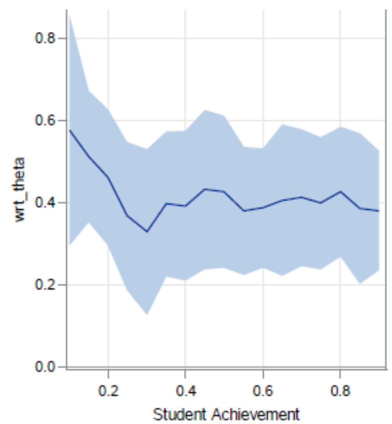


Grade 10

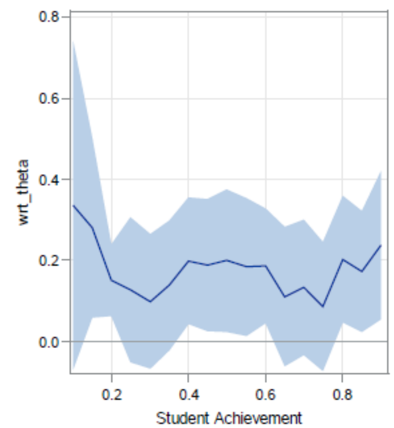
Appendix E: Quantile Correlations between FAIR-FS Word Recognition and TOWRE Real Word



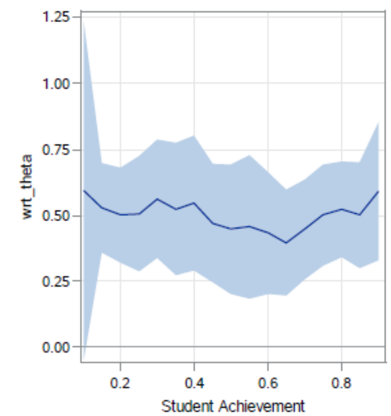
Grade 3



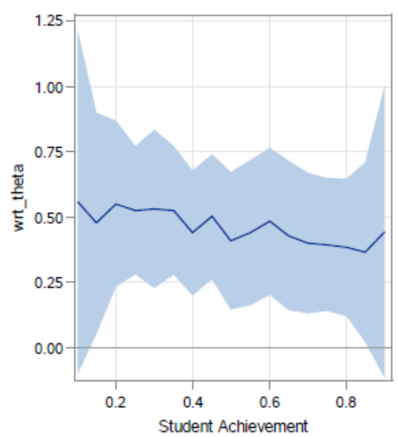
Grade 4



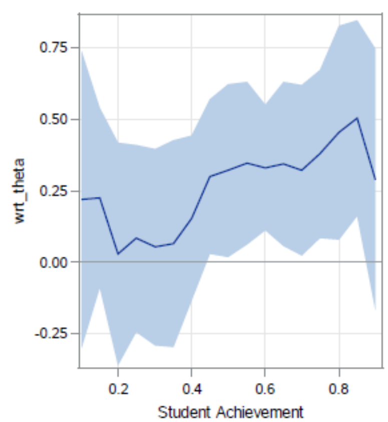
Grade 5



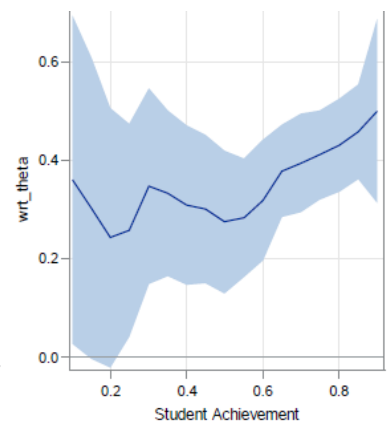
Grade 6



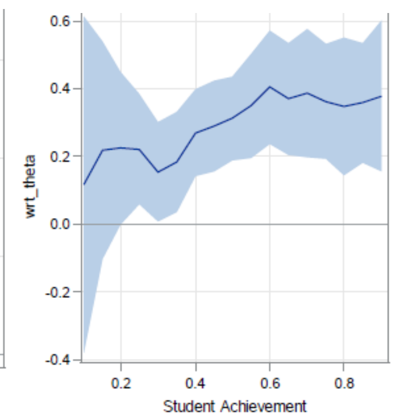
Grade 7



Grade 8

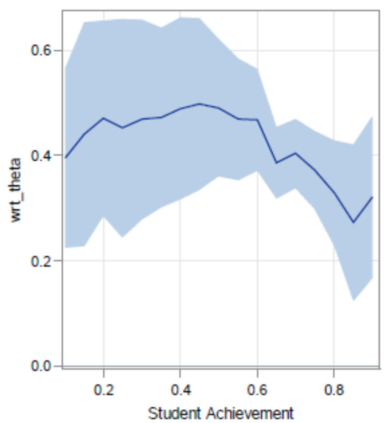


Grade 9

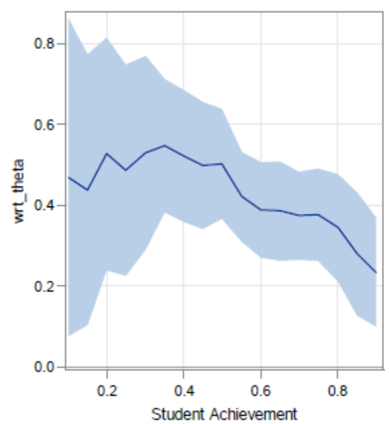


Grade 10

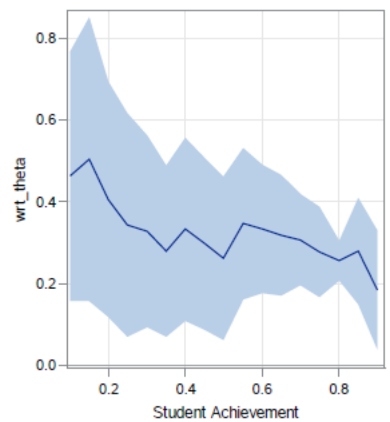
Appendix F: Quantile Correlations between FAIR-FS Word Recognition and TOWRE Non-Word



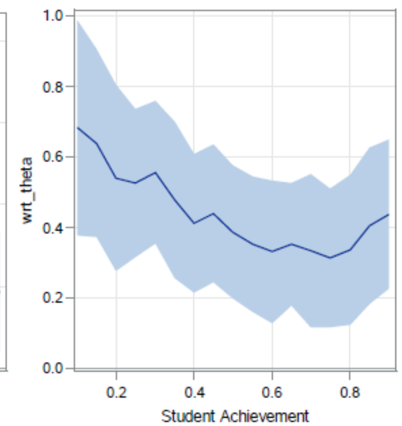
Grade 3



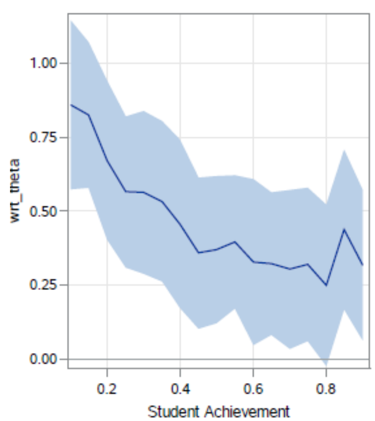
Grade 4



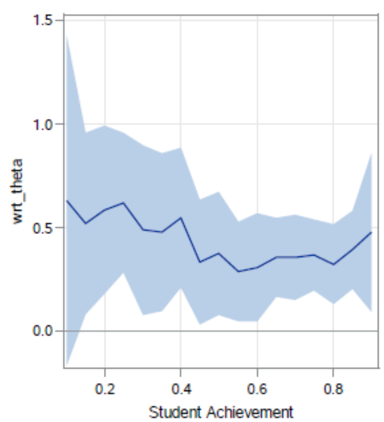
Grade 5



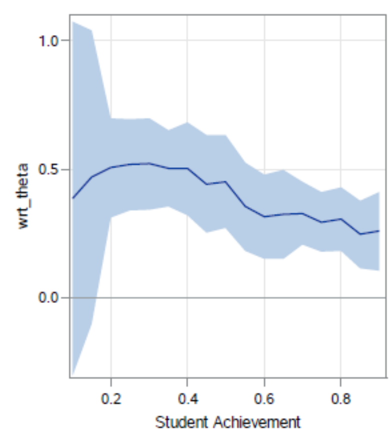
Grade 6



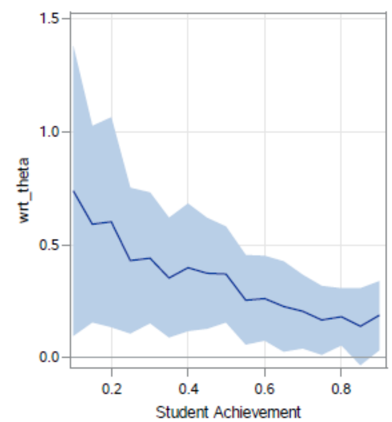
Grade 7



Grade 8

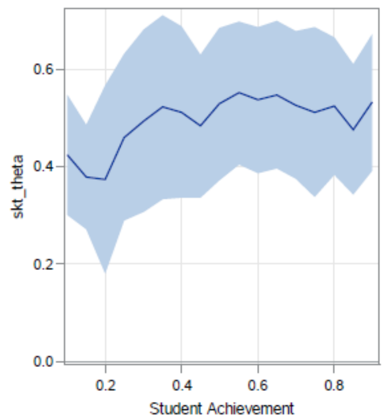


Grade 9

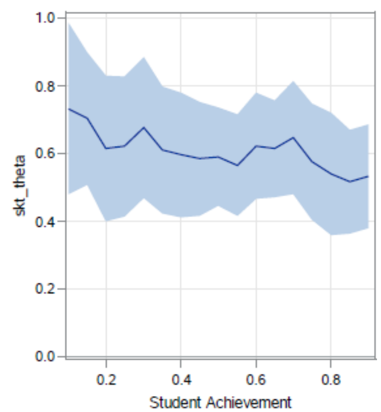


Grade 10

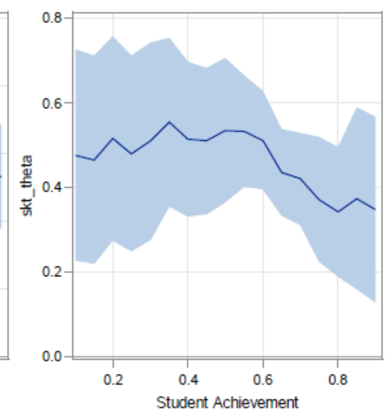
Appendix G: Quantile Correlations between FAIR-FS Syntax Knowledge and GJT



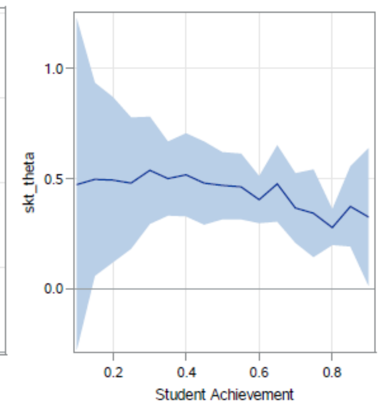
Grade 3



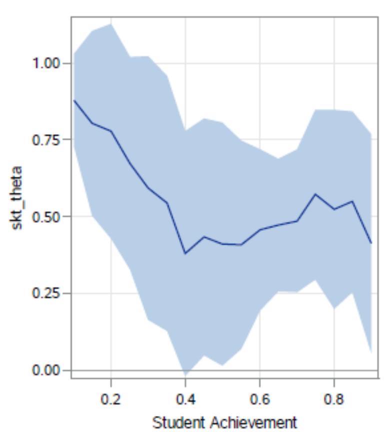
Grade 4



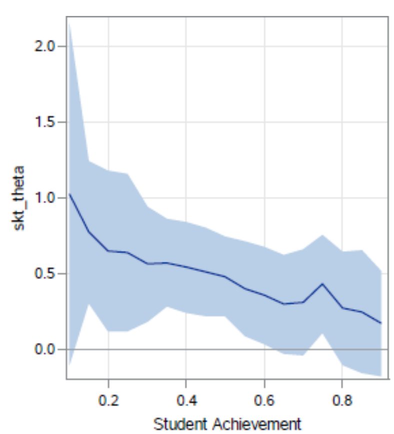
Grade 5



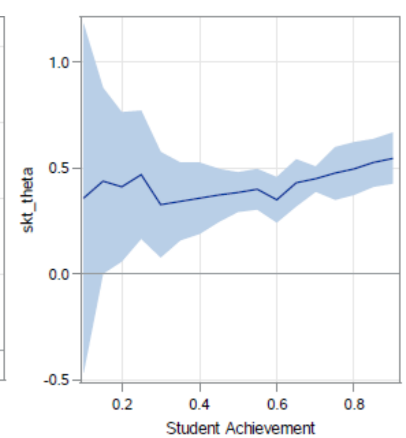
Grade 6



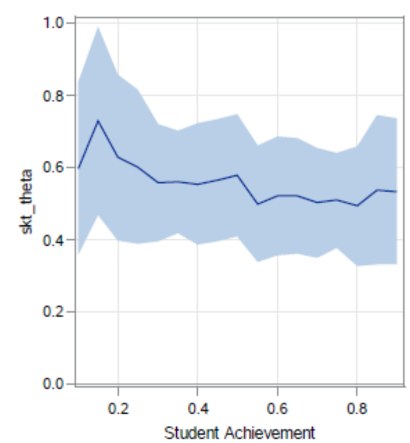
Grade 7



Grade 8



Grade 9



Grade 10